UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Trabajo fin de Máster en Ingeniería y ciencia de datos

Análisis estadístico de indicadores de juventud en espectros RP de estrellas enanas ultrafrías de Gaia

Autor
Daniel Diez Tabara

Dirigido por
Luis Sarro Baro

Madrid
Febrero 2023

Resumen

Durante la misión espacial del telescopio Gaia de la Agencia Espacial Europea se obtendrán una gran cantidad de datos de objetos celestes. Entre esa gran cantidad de datos se encuentran medidas astrométricas (como distancias, posiciones o movimientos propios) y también medidas fotométricas (como brillos aparentes y espectros). Estas medidas fotométricas forman parte del conjunto de datos que periódicamente la sonda envía a la base en la tierra y sobre las que se centra este trabajo. En concreto, se analizará los espectros RP (longitudes de onda características de la luz roja 640-1050 nm) de un subconjunto de estrellas llamadas enanas ultrafrías y se aplicarán diversas técnicas de creación de espacios latentes de menor dimensión, para detectar indicios de juventud en dichos espectros.

Índice general

Índice de figuras

Índice de cuadros

Acrónimos

PCA	Principal Component Analysis
SVD	Singular Value Decomposition
GDR3	Gaia Data Release 3
CCD	Charger Coupled Devices
SDSS	Sloan Digital Sky Survey

Capítulo 1

Introducción

1.1. Contexto

La sonda Gaia fue lanzada en diciembre de 2013 y su objetivo principal es recoger los datos astrométricos de nuestra galaxia para poder crear el mapa de la Vía Láctea más preciso hasta la fecha [ESA22b]. La sonda cuenta con dos telescopios que observan el espacio al mismo tiempo y que envían simultáneamente la luz a una cámara formada por 938 millones de píxeles compuesta por 106 detectores CCD (Charger Coupled Devices). Cuando algún objeto celeste se selecciona por la unidad de procesamiento para ser observado en detalle, algunas decenas de píxeles de cada CCD almacenan la información fotométrica de dicho objeto. El resto de CCD se dedican al objetivo principal de Gaia, que consiste en el análisis astrométrico de la vía Láctea.

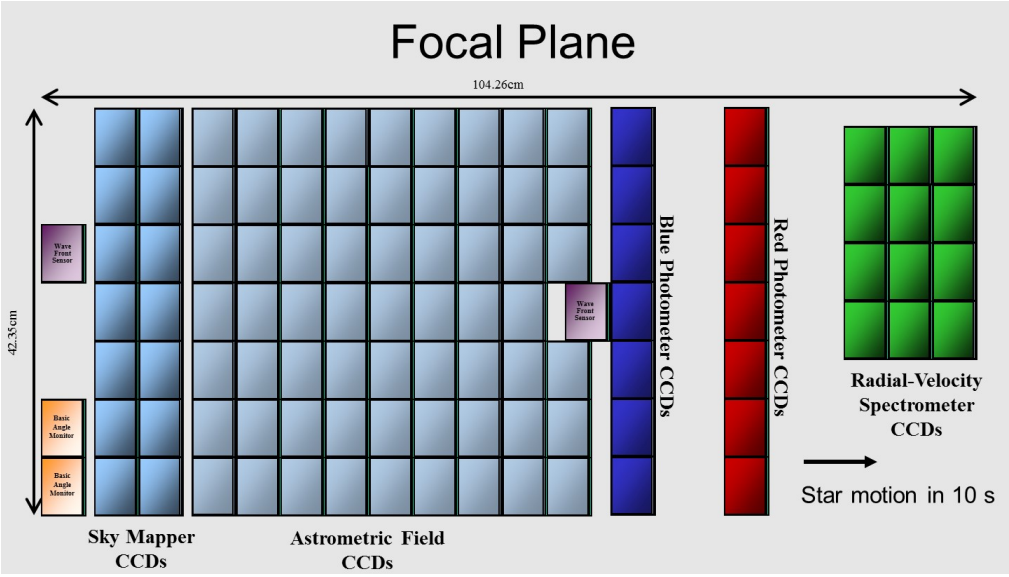

Figura 1.1: Esquema de la composición del panel de CCD de Gaia. Las hileras verticales *Blue Photometer CCDs* y *Red Photometer CCDs* son las encargadas de registrar y procesar la información espectral de los objetos celestes. El resto de CCDs se dedican al objetivo principal de Gaia de captura de información astrométrica. Dos prismas son los encargados de dispersar la luz *azul* y *roja* para que sea captada por los CCD dedicados.

Los instrumentos fotométricos miden la distribución de energía espectral de todos los objetos seleccionados en el mismo momento que se toman los datos astrométricos. Los espectrógrafos se apoyan en dos prismas que dispersan la luz incidente, uno de ellos se centra en la dispersión de la luz azul (llamado BP) operando entre 330 nm y 680 nm y el otro se centra en la dispersión de la luz roja (llamado RP) operando entre 640 nm y 1050 nm (figura 1.1). Ambos espectrómetros cuentan con 7 CCD [Col16]. Como resultado se tiene que a la hora se observan dos millones de objetos celestes y se genera un total de 50 gigabytes al día.

El 13 de junio de 2022 se publicó el llamado Gaia Data Release 3 (GDR3) con información astrométrica y fotométrica de 1.800 millones de objetos celestes y el detalle de los espectros de casi 220 millones de estos objetos [ESA22a].

Se estima que las enanas ultrafrías representan un gran porcentaje de los objetos astronómicos de la Vía Láctea. Aproximadamente el 15 % de los objetos estelares que son detectados en los sistemas estelares cercanos pertenecen a esta categoría. Además, se considera que, debido a la baja cantidad de masa que hay en la zona de nacimiento de este tipo de objetos, hace que los discos protoplanetarios tengan también baja masa y que sea entornos propicios para la creación de numerosos planetas pequeños [CPG05]. Por eso que son un foco de investigación en los proyectos de búsqueda de vida extraterrestre [Sho16]. No así tanto los entornos donde los discos protoplanetarios son muy masivos, que son propensos a crear grandes estrellas con un número reducido de planetas gigantes.

Se denomina enana ultrafría a los objetos estelares o subestelares cuya temperatura superficial es inferior a los 2700 Kelvin y con tipo espectral M7 o más tardío. Son subestelares aquellos objetos que no consiguen iniciar o mantener reacciones nucleares de forma continua por la falta de masa, a diferencia de lo que sucede con las estrellas que se encuentran en la secuencia principal. Esta nueva categoría de estrella fue introducida en 1997 por J. Davy Kirkpatrick [JDKI97]. En este trabajo se utilizará el término "estrella" por simplicidad, aunque la categoría de enana ultrafría incluye tanto estrellas como enanas marrones. Se estima su vida en cientos de miles de millones de años o incluso billones de años. Se pueden estudiar este tipo de objetos como el enlace entre planetas gigantes y estrellas poco masivas. Estos objetos emiten su flujo electromagnético mayoritariamente en el infrarrojo, extendiéndose algo de energía al rojo visible, es por eso que se analizará el espectro RP recogido por la sonda Gaia.

Como son estos objetos tan particulares y aparentemente distintos de las estrellas comunes que se encuentran en la secuencia principal, ¿Podrán encontrarse indicios de la edad bajo las características de sus espectros? Para responder a esta pregunta, se usarán tres técnicas de creación de espacios latentes de menor dimensión sobre los espectros. Se analizará el efecto que sobre estos nuevos espacios tiene la edad de los objetos. En concreto, se analizará un conjunto de datos formado por 27.500 objetos ultrafríos, de los que se saben cuáles están en la secuencia principal y cuáles son estrellas de una edad temprana.

Existen numerosas técnicas para la creación de espacios latentes de menos dimensión (reducción de dimensionalidad), pero en este trabajo se pondrán en práctica tres de ellas por los resultados que han dado en otros trabajos sobre el análisis de espectros fotométricos. En concreto, se podrán bajo objeto de estudio las técnicas: Principal Componet Analysis, Autoencoders y Mapas de Difusión.

1.2. Estructura de la memoria

Esta memoria se compone de seis secciones:

- La primera es la introducción, a la que pertenece este apartado.
- La segunda sección se dedica al estado del arte, donde se revisarán las últimas aplicaciones de las técnicas de extracción de características principales.
- En la tercera sección, se entrará a realizar una breve descripción de los datos usados en este trabajo.
- Una vez descritos los datos, en la cuarta sección se expondrá la metodología del trabajo, que se compone de cuatro partes:
  - Preprocesado de datos
  - Técnicas de creación de espacios latentes
  - Visualización de los resultados
  - Validaciones.
- En la quinta sección se expondrán los resultados obtenidos.
- La última sección se dedica a las conclusiones y trabajos futuros.

Capítulo 2

Estado del arte

El análisis de datos de alta dimensión es un problema importante en muchos ámbitos de la ciencia y como no, en el de la astrofísica. Los métodos de reducción de dimensionalidad permiten extraer de un conjunto de datos de alta dimensión (con un gran número de variables) las características principales, convirtiendo el conjunto de datos en un nuevo conjunto con muchas menos dimensiones (muchas menos variables) que contendrán las la esencia principal, es decir, aquellos que contienen la información importante para el objeto en estudio. Al nuevo espacio formado se le llama comúnmente espacio latente. Un ámbito donde se utiliza habitualmente este tipo de análisis es la exploración visual de información, donde reducir las dimensiones a dos o tres variables permite al ojo humano analizar e interpretar visualmente la información, cosa que con, por ejemplo, 100 dimensiones es imposible. El proceso de búsqueda del espacio latente se basa en proyectar los datos n-dimensionales de entrada en nuevos espacios l-dimensionales de salida (siendo n>l) tratando siempre de mantener la mayor cantidad de información posible en cada una de las proyecciones. La información que no se pueda mantener en cada proyección se pierde, eliminado así las características que no contienen información importante. Las primeras referencias sobre este tipo de técnicas se remontan hasta 1901 desarrolladas por Pearson y publicadas en Philosophical Magazine [Pea01]. Posteriormente, en la década de 1930, Hotelling en Journal of Educational Psychology [Hot33] plantea la idea de crear un nuevo espacio que sea combinación lineal del espacio original, pero teniendo en especial consideración aquellas variables que presenten coeficientes de relación más grandes (en valor absoluto).

Desde la aparición de las primeras publicaciones hasta el día de hoy, estas técnicas han tenido un largo recorrido teórico y práctico, lo que ha dado lugar a una gran cantidad de variedades. Todas ellas mantienen el objetivo de conservar la esencia de los datos originales en el nuevo espacio latente, para que los procesos y análisis posteriores no se vean afectados por la posible pérdida de información. Existen varios criterios para la clasificación de dichas técnicas, la forma más común de clasificarlas son las que distinguen entre la forma fundamental que tienen de obtener el espacio latente. Estas se dividen en técnicas Lineales y No Lineales. Los métodos Lineales son los más simples, usados y conocidos, como por ejemplo el tradicional PCA. Estos se fundamentan en la realización de transformaciones lineales de los espacios formados por las variables originales para buscar los nuevos espacios. Una trasformación F es lineal que transforma

el espacio V en W, si y solo si cumple con:

$$F(u+v) = F(u) + F(v) \qquad \forall u, v \in V \tag{2.1}$$

$$F(ku) = kF(u) \qquad \forall u \in V, \forall k \in \mathcal{R} \tag{2.2}$$

Por el contrario, los métodos no lineales son aquellos que buscando los nuevos espacios latentes no cumplen con la definición de la transformación lineal, ni por lo tanto, aprovechan sus propiedades. La utilización de estos suele estar indicados en aquellos casos donde se sabe o intuye que los datos se encuentran localizados un espacio no lineal. Bajo esta definición tan amplia caben un gran número de métodos, algunos de ellos basado en algoritmos de aprendizaje automático como *Autoencoders* u otros que usan funciones no lineales (normalmente llamadas kernel) para hacer la transformación entre espacios como, por ejemplo, *Mapas de Difusión*.

A continuación, en la figura 2.1, se presenta un esquema con distintas técnicas para la búsqueda de los espacios latentes basándonos en esta clasificación.

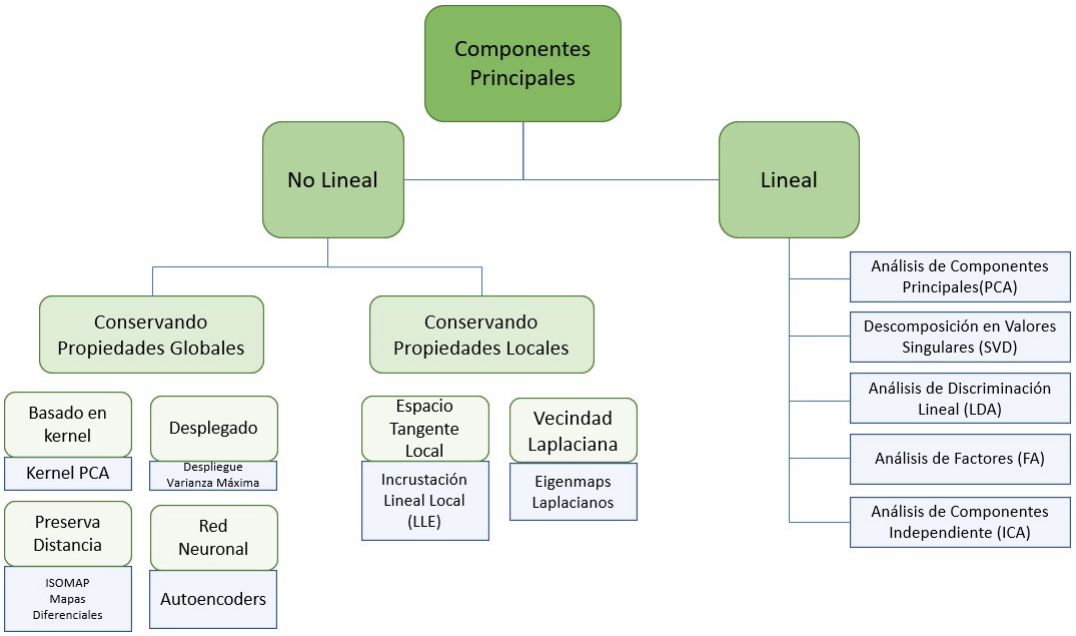

Figura 2.1: Tipos de técnicas de búsqueda de espacios latentes. Las dos grandes *familias* basadas en la *linealidad* del método de transformación utilizado para la búsqueda de los nuevos espacios latentes. Las técnicas lineales están recomendadas en los casos en los que los datos están habitando espacios lineales. Cuando se sabe o intuye que el espacio no es lineal, las técnicas no lineales son las indicadas. La selección de la técnica apropiada depende mucho del objetivo perseguido. Por ejemplo, si lo que se pretende es dar relevancia o peso a la localidad de los datos, deberán seleccionarse técnicas que se preocupen de conservar las propiedades locales. Para este trabajo se va a usar Análisis de Componentes Principales, Autoencoders y Mapas de Difusión. La primera se tomará a modo de referencia.

2.1. Aplicaciones PCA

Por ser PCA una de las técnicas más usadas, hoy en día se sigue investigando, adaptando y evolucionando esta para su aplicación en casi todas las ramas de la ciencia. Por ejemplo, existen métodos adaptados en los que la selección de los criterios que definen al nuevo espacio latente, y por tanto, a las características principales, se basa en seleccionar los autovalores que maximizan la curtosis (cuanto mayor es la curtosis de una variable aleatoria, mayor es la concentración de valores entorno a la media). Estos métodos son llamados Cokurtosis-PCA. Algunas de sus aplicaciones se presentan, por ejemplo, en el estudio del comportamiento físico de fluidos [JKR+22].

Otra modificación que se ha aplicado al tradicional PCA consiste en la aplicación de la ley de distribución de probabilidad de Marchenko-Pastur para separar los valores propios que se relacionan con información aleatoria (ruido) en la matriz de covarianza, de aquellos que representan la información intrínseca de la señal. Bajo el teorema de Marchenko-Pastur [MP67] si los elementos de una matriz son aleatorios, sus valores propios siguen una distribución específica llamada Marchenko-Pastur, entonces aquellos valores que no sigan esa distribución contendrán la información de la parte no aleatoria, es decir, de la señal. Esta aplicación es muy útil en los procesos de eliminación del ruido en la matriz de correlaciones y para extraer la información relevante. Un ámbito en el que se usa con buenos resultados esta técnica es en el de la radiología en los procesos de limpieza y optimización de la calidad de las imágenes de resonancia magnética. Son bastante útiles para el uso en la fase de diagnóstico y seguimientos de afecciones tumorales en cerebros [AALV+21].

La aplicabilidad de esta técnica sobre el campo de la astrofísica también resulta muy fructífera. En el caso de estudio de este trabajo se aplicará sobre los espectros RP publicados por Gaia en la GDR3, pero existes otros trabajos recientes donde se aplica, por ejemplo, para profundizar en el conocimiento sobre los campos magnéticos que producen y rodean a las estrellas. Un reciente trabajo [LD22] procede a recuperar las principales características de las topologías magnéticas a gran escala de las estrellas, comparando su sencillez y rendimiento con otras técnicas de análisis de información tradicionales en el campo que son mucho más tediosas, complejas y en algunos casos que necesitan que los investigadores tengan conocimientos muy profundos en dicha materia.

También otros trabajos usan esta herramienta en la tarea de preprocesado de datos, por ejemplo cuando a través de los espectros se intentan construir modelos que permitan encontrar estimadores de la relación que hay entre la masa y la luz estelar de galaxias [PTC+19]. En este caso se usa para obtener las seis componentes principales ocultas de una biblioteca de espectros sintéticos de 40.000 estrellas. O por ejemplo, muy relacionado con el uso que se hace en el ámbito de la eliminación del ruido en imágenes de resonancias magnéticas, se encuentran casos de uso directamente relacionados con la separación de las características de las fluctuaciones en el espectro derivadas de procesos físicos internos de las estrellas, de las características derivadas de las atenuaciones que sobre el espectro provocan el tránsito de planetas en sus órbitas alrededor de las

estrellas[DCD+17].

2.2. Aplicaciones Autoencoders

Los Autoencoders pertenecen a la familia de las redes neuronales y están diseñados para aprender las características internas de conjuntos de datos, maximizando el aprendizaje para luego ser capaces de sintetizar nuevos conjuntos de datos que se asemejen al conjunto de datos original. Se puede utilizar únicamente el proceso de aprendizaje de las características de los datos para crear así las bases del nuevo espacio latente, aunque no exista la necesidad de sintetizar nuevos conjuntos de datos. Las primeras referencias y aplicaciones de este tipo de red neuronal se remontan a la década de 1980 [DH85] donde el objetivo mayoritario de su uso era precisamente la extracción de componentes principales. No obstante, en los últimos tiempos están ganando fama por su éxito como modelos generativos, es decir, para producir nuevos conjuntos de datos, como sintetizadores de imágenes hiperrealistas.

El uso en diferentes ramas de la ciencia es exitoso. De nuevo, en el campo de la medicina aplicada, en la lucha contra enfermedades degenerativas, existen casos de uso donde utilizan este tipo de redes neuronales complementadas con redes adversarias generativas, para crear nuevos modelos que ayuden a identificar los tipos tan heterogéneos de disfunciones cerebrales que produce la enfermedad del Alzheimer [WZZ+22]. Otros caso de uso lejos de la medicina se pueden encontrar en [TJI+22] donde los autores combinan estos algoritmos junto con técnicas de visión artificial para descubrir patrones desconocidos y relevantes en la evolución de las nubes, reduciendo así la incertidumbre en las simulaciones y, por lo tanto, en las predicciones.

Por otro lado, en el ámbito de la cosmología existen aplicaciones recientes alrededor de la detección de ondas gravitacionales. Algunos casos muy específicos centrados en la eliminación del ruido de las señales emitidas por sistemas binarios de agujeros negros en rotación. En este trabajo [BTB22] se reconstruye la señal original libre de ruido usando este tipo de redes neuronales. Aunque no se usa explícitamente el espacio latente producido por el codificador, es esencial para que el decodificador pueda eliminar el ruido de la señal original. También algunos trabajos usan variantes de este tipo de redes neuronales para mejorar el proceso de detección de ondas gravitacionales. En [MVS+21] los autores analizan la bondad de Autoencoders recurrentes comparados con las arquitecturas tradicionales.

Por último, más cerca del ámbito de este trabajo y relacionado con los espectros de objetos astronómicos, se encuentra un claro ejemplo de búsqueda del espacio latente con Autoencoders. En este caso [PPVC21] los autores utilizan los espectros publicados por el proyecto SDSS `https://www.sdss.org/`para extraerles las características principales utilizando Autoencoders del tipo Variacionales intentando capturar las relaciones no lineales que puedan existir entre los espectros. Concluyen que estas técnicas producen mejores resultados que la tradicional técnica Principal Component Análisis.

2.3. Aplicaciones Mapas de Difusión

Este método de búsqueda de nuevos espacios latentes fue introducido por Ronald R. Coifman y S. Lafon en 2005 [CR05] cuyo detalle matemático se dará en las siguientes secciones. Se basa en la idea de que las características más relevantes de un conjunto de datos se pueden capturar mediante la construcción de un grafo, en el que los nodos representan los datos y las aristas representan las relaciones o similitudes entre ellos. A medida que se recorre el grafo, los datos que están más cerca en el espacio de características originales también estarán más cerca en el grafo. De esta manera, los Mapas de Difusión pueden ayudar a visualizar y comprender mejor la estructura subyacente de los datos en un espacio de dimensionalidad reducida. Se comienza calculando una matriz de similitud entre los datos a partir de algún tipo de medida de distancia, como la distancia euclídea. A continuación, se aplica un proceso de "difusión" a esta matriz, que tiende a conectar los datos que están más cerca entre sí en el espacio original. Esto se puede hacer mediante el uso de una matriz de transición que se aplica iterativamente a la matriz de similitud, enfatizando cada vez más las conexiones entre los datos más similares. Un enfoque común es utilizar el operador de difusión de Markov, que propaga las similitudes a lo largo del grafo mediante un proceso iterativo. Al final del proceso, se obtiene una representación de los datos en un espacio de dimensionalidad reducida que mantiene las propiedades de similitud de los datos originales. Ahora cada observación es determinada en el nuevo espacio de difusión a través de las llamadas coordenadas de difusión.

A modo de resumen, el método consiste en obtener los l autovectores propios de la matriz de probabilidad formada por las probabilidades de salto entre los puntos (siendo los puntos cada una de las observaciones en el espacio formado por las m variables originales y siendo $m > l$). La probabilidad se relaciona a través de una función no lineal con la distancia entre puntos, de forma que el método conserva la esencia de las relaciones de vecindad entre los puntos.

La búsqueda de nuevos espacios latentes con Mapas de Difusión se aplica en diversas áreas de la ciencia. Algunas de las últimas aplicaciones están relacionadas con los estudios de pérdida de biodiversidad debido a los efectos del cambio climático. Gracias a esta técnica se permite la identificación de características funcionales en las especies, para poder cuantificarlas y así poder ver el impacto del cambio climático sobre la diversidad funcional. El trabajo de Alexey Ryabov del 2021 se analiza los resultados de esta técnica sobre una cierta especie de fitoplancton del mar Báltico [RBH+21]. Algunos trabajos más antiguos y muy interesantes sobre la aplicabilidad de dicha técnica consisten en solucionar el problema de la localización física de dispositivos que están conectados a una red inalámbrica. En el trabajo de Amin Ghafourian y Orestis Georgiou en 2021 [GGBG19] muestran que esta técnica devuelve un resultado muy preciso a la hora de localizar los dispositivos dentro de la red. En algunos casos reduce el coste computacional, comparado con algunas de las técnicas de localización tradicionales, incluso son más resistentes a la interferencia que el ruido introduce a la hora de realizar las localizaciones.

Desde el punto de vista de aplicabilidad en el campo de la astrofísica, esta técnica también está dando sus frutos. Un trabajo publicado en 2021 analiza, con la ayuda de este método, la estabilidad a largo plazo de algunos objetos celestes del sistema solar externo. A partir de un análisis de frecuencia de las perturbaciones de los planetas gigantes y de 34 objetos situados más allá de la órbita de Neptuno, simulan 200 escenarios para los próximos 1.000 millones de años y transforman esos datos a un nuevo espacio reducido en dimensiones aplicando Mapas de Difusión para facilitar el análisis numérico, encontrando así qué objetos saldrán del sistema solar, cuáles caerán al interior y cuáles se pueden considerar moviéndose en órbitas estables [MGPLW21]. También es aplicado para encontrar las zonas de estabilidad de los anillos de Neptuno. Aplicando Mapas de difusión, se puede hacer una representación más simplista (gracias a las coordenadas de difusión en el nuevo espacio latente de dimensión inferior) y, por lo tanto, facilitar el análisis sobre la estabilidad de cada uno de los anillos [GWMMG20].

Más relacionado con los análisis fotométricos de objetos celestes, se pueden encontrar trabajos en los que se proponen los Mapas de Difusión en las tareas de preprocesado de datos dentro de tareas de procesos más complejos. Se encuentra el trabajo de Joseph W. Richards y Darren Homrighausen de 2012 [SP12] en el que se propone un método para tipificar los espectros de las estrellas supernovas. En primera instancia se aplica una técnica de reducción de dimensionalidad basada en Mapas de Difusión, para eliminar características subyacentes de los datos no deseadas. A continuación, con el espacio de datos reducido resultante se procede a entrenar un modelo de clasificación basado en bosques aleatorios.

Capítulo 3

Datos

Según la información publicada en la página web de la Agencia Espacial Europea sobre los datos contenidos en el llamado Gaia Data Release 3 y accesible desde https://www.cosmos.esa.int/web/GAIA/dr3, el conjunto de datos que se pone a disposición del público consiste en 1.811 millones de objetos celestes con información astrométrica y 211 millones de objetos con los espectros RP/BP. Siendo así la primera vez que la misión publica datos sobre los espectros. Para este trabajo se cuenta con información astrométrica y fotométrica de 27.478 objetos celestes. A priori se sabe que 2.668 objetos son clasificados como supuestamente jóvenes y los 24.810 están en la llamada secuencia principal, parte del ciclo de vida de las estrellas en el que permanecen prácticamente durante toda su existencia. Para las estrellas ultrafrías, la secuencia principal comienza al final de la fase de contracción inicial (para la gran mayoría de tipos de estrellas, la secuencia principal comienzan cuando se producen reacciones nucleares de forma continuada). El conjunto de datos contiene un identificador único para cada registro y 148 variables. De estas, 19 variables pertenecen al conjunto de medidas astrométricas, 9 corresponden con información fotométrica y 120 se reservan para el espectro RP.

3.1. Datos Astrométricos

Aunque no tiene relevancia para este trabajo por no usarse, se pasan a describir a alto nivel la información astrométrica que contiene el conjunto de datos:

- Se reservan 4 variables para identificar la posición en el cielo con su correspondiente margen de error.
- Se incluyen 4 variables para indicar el movimiento aparente en el cielo con sus correspondientes márgenes de error.
- Con 1 variable se indica el número de observaciones usadas para la recogida de la información.
- Para indicar la bondad de los ajustes de los modelos que se han usado para las medidas astrométricas se proveen de 4 variables.

- Se usan 3 variables para dar información sobre los parámetros usados para la determinación de las imágenes (simetría en las imágenes).

- Por último, se indica la temperatura efectiva estimada del objeto celeste. Dentro del catálogo GDR3 el objetivo de esta última variable es identificar las estrellas enanas ultrafrías. El rango de temperatura de los datos proporcionados en secuencia principal varía desde 1132 K hasta 2700 K y el rango para las estrellas clasificadas como jóvenes varía desde los 2155 K hasta los 2700 K.

3.2. Datos Fotométricos

Las variables fotométricas tampoco se usan para este trabajo, pero el contenido en el conjunto de datos a alto nivel es:

- Se usan 3 para las medias de magnitudes en las bandas G, G_{bp} y G_{rp}

- Otras 3 para la media el flujo de las bandas G, G_{bp} y G_{rp}

- Y por último 3 más para los errores de las medidas del flujo de las bandas G, G_{bp} y G_{rp}

3.3. Espectro

Los espectros de luz de las estrellas son una herramienta muy útil para entender la química y la física de la superficie de las estrellas. Estos espectros muestran la cantidad de energía electromagnética que una estrella emite a diferentes longitudes de onda. Normalmente estos espectros están compuestos por una serie de líneas de absorción y emisión, que se forman cuando la luz interactúa con los átomos en la superficie estelar. Estas líneas muestran la composición química de la superficie estelar, por lo que los astrónomos pueden utilizar esta información para entender mejor la física de la estrella. Las líneas de emisión también pueden usarse para determinar la temperatura superficial de la estrella.

Los espectros se pueden usar para estudiar la evolución estelar. Las líneas de absorción y emisión diferentes, indican diferentes etapas de la evolución de una estrella. Por ejemplo, los espectros de luz pueden mostrar la edad de una estrella a partir de la cantidad de ciertos elementos químicos presentes en su superficie. Los espectros también revelan el tamaño, la masa y la luminosidad de una estrella, permitiendo a los astrónomos determinar la evolución de la estrella a través del tiempo.

En resumen, la forma de los espectros está mostrando información sobre su composición, temperatura y edad de las estrellas.

El espectro utilizado en este trabajo viene caracterizado por un total de 120 variables. Estas tienen su origen en el flujo de fotones recibido en cada píxel de cada CCD reservado para las longitudes de ondas en el rango RP. Las unidades iniciales de estas variables son $e^{-}seudopixel/s$, es decir, número de electrones recibidos por seudopíxel en cada segundo de observación del objeto celeste. Estas 120 variables proporcionadas para este trabajo han sufrido un proceso de normalización (el área contenida debajo de cada curva es igual a la unidad), por lo que carecen de unidades. El objetivo de este proceso de normalización, ajeno a este trabajo, tiene su naturaleza en tratar de poner a todas las medidas de todas las estrellas en la misma escala. Es decir, para no tener en cuenta el efecto que la distancia pueda tener sobre la cantidad de electrones recibidos ni, por lo tanto, el brillo aparente de cada estrella. Tras este proceso de normalización, dos estrellas similares que pudieran estar a distancias muy diferentes, producirán espectros similares, independientemente de que se reciban más electrones por segundo de la que esté más cercana a la sonda.

Los flujos normalizados de electrones recibidos corresponden con las longitudes de onda del espectro RP, que van desde los 640 hasta los 1050 nanómetros (10^{-9}metros). En el conjunto de datos se identifican como $flux1, flux2...flux120$.

Morfología

En la figura 3.1 se muestran las medias de los espectros de cada tipo de estrella por rangos de temperatura. Cada rango contempla 100 K, empezando en los 1100 K y terminando en los 2700 K. Se puede observar que de forma generalizada los espectros de las estrellas jóvenes tienen picos menos acentuados que los espectros de las estrellas en secuencia principal. Como los espectros han sido normalizados, puede haber longitudes de onda donde el valor del flujo sea superior para un tipo de estrella, pero eso implicará que en otras longitudes de onda ese tipo de estrellas tendrá valores inferiores del flujo. Eso se traduce en que los espectros de ambos tipos de estrellas se irán cruzando a lo largo de la representación. En general se pueden observar claramente 3 picos de intensidad con sus 2 correspondientes valles. Cuánto más frías son las estrellas, los picos son menos acentuados y casi inapreciables para las estrellas jóvenes. Para estas estrellas más frías, la morfología de los espectros son muy diferentes. Por el contrario, cuanto más alta empieza a ser la temperatura de la estrella, empiezan a acentuarse los picos de las estrellas jóvenes, hasta llegar un punto (para temperaturas en el rango 2500-2700 K) que la morfología de los espectros de ambos tipos de estrellas son muy parecidos.

Analizando un poco más en detalle los espectros, se procede a representar la diferencia del valor medio de los flujos y la diferencia de curvatura de los espectros, de cada tipo de estrella, frente a la longitud de onda. Se pretende analizar como es la diferencia de curvatura de los espectros para cada tipo de estrella. El estudio de la curvatura de una curva, permite obtener la tasa de variación de la dirección de una curva con respecto a su longitud. Para el cálculo de la curvatura del espectro es necesario realizar el cálculo de la primera y segunda derivada de forma numérica con la función numpy.gradient().

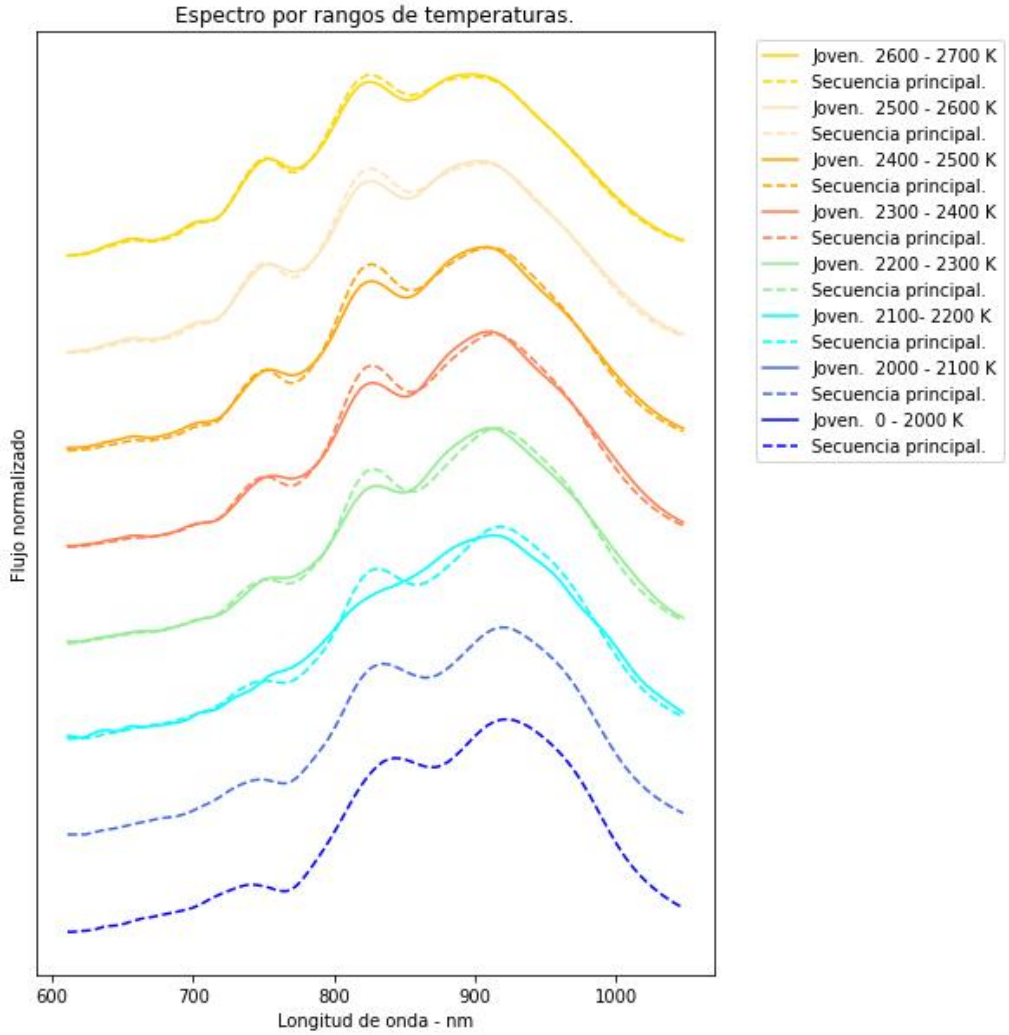

Figura 3.1: Representación de los espectros por rango de temperatura. Se muestra el espectro medio para las estrellas en los siguientes rangos de temperaturas: 2600K - 2700K, 2500K-2600K, 2400K-2500K,2300K-2400K,2200K-2300K,2100K-2200K,2000K-2100K y menores de 2000K. Con el trazado en línea continua para las estrellas jóvenes y con el trazado discontinuo para las estrellas en secuencia principal. Se puede observar un perfil común para todos los espectros.

La curvatura es calculada a partir de la siguiente ecuación (no es objeto de este trabajo el desarrollo de la misma):

$$\kappa = \frac{\|x''y' - x'y''\|}{((x')^2 + (y')^2)^{3/2}} \tag{3.1}$$

Se puede observar en la figura 3.2 como la diferencia de la media de los espectros y la diferencia del valor de la curvatura de ambos tipos de estrella, disminuyen según aumenta la temperatura de las estrellas. En las estrellas en rango de temperaturas de 2100 - 2200 K existen varias longitudes de onda que presentan una curvatura importante, destacando las que se tienen entre 750 y 900 nm. Conforme va aumentando la temperatura de las estrellas, la amplitud de los picos va disminuyendo (las curvaturas de las estrellas son más similares), hasta que prácticamente solo queda un pico pronunciado alrededor de los 775 nm.

¿Serán suficientemente significativas estas diferencias morfológicas para que sean evidenciadas en los nuevos espacios latentes? ¿Recogerán las nuevas variables estos indicios de juventud de las estrellas? ¿Se podrá verificar que las nuevas variables estarán influidas por la temperatura de la estrella, como se indicaba en la introducción de esta sección?

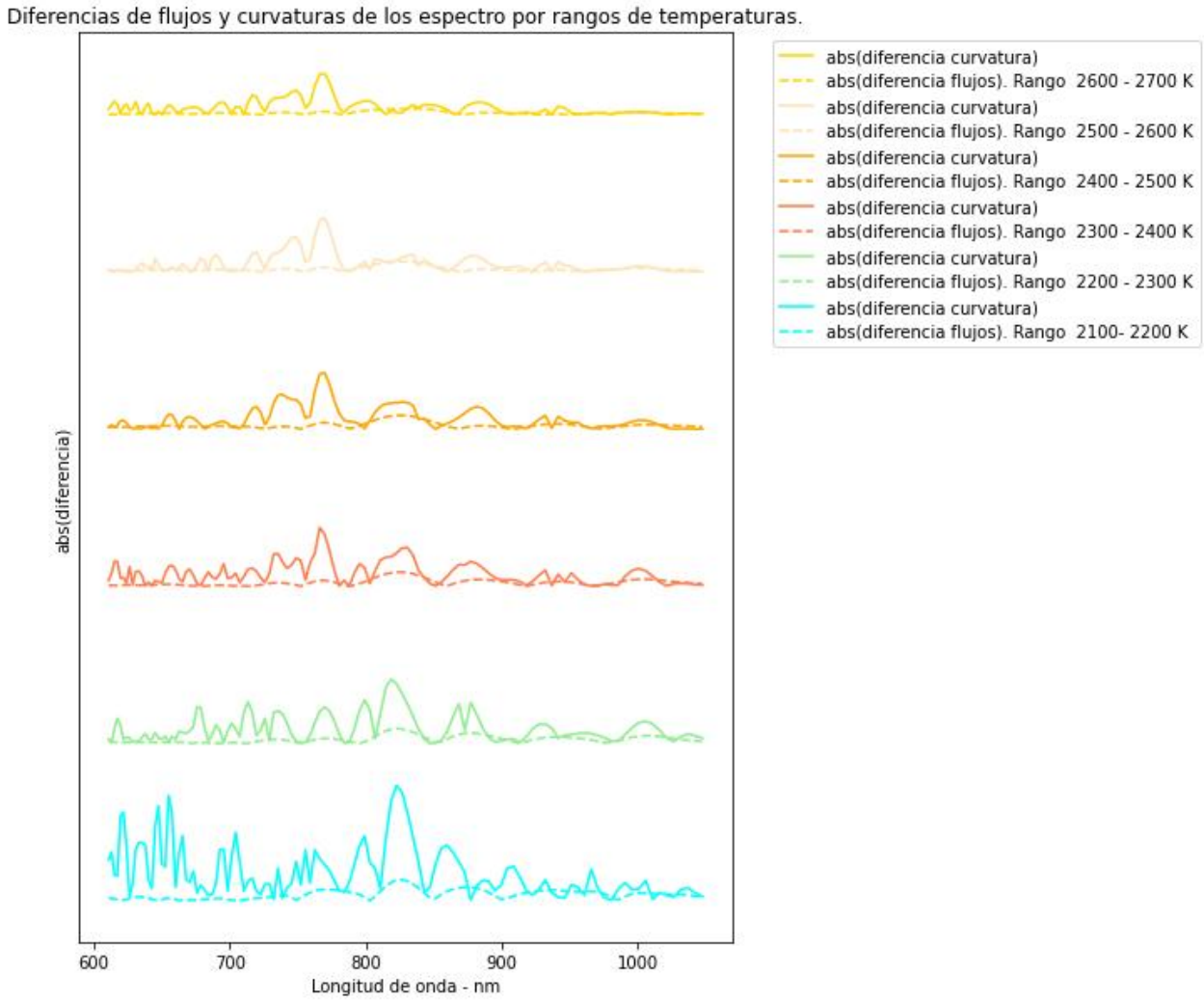

Figura 3.2: Representación de la diferencia de los valores y de la diferencia de curvatura de los espectros entre los dos tipos de estrellas. Se muestran separadas en los siguientes rangos de temperaturas: 2600 - 2700 K, 2500 - 2600 K, 2400 - 2500 K, 2300 - 2400 K, 2200 - 2300 K, 2100 - 2200 K, 2000 - 2100K y menores de 2000 K. Con el trazado en línea continua se identifica la diferencia en valor absoluto de los espectros y con línea continua la diferencia de curvatura. Conforme aumenta la temperatura de las estrellas, van disminuyendo ambas diferencias.

Estadísticas

Se realiza un análisis de las variables estadísticas más significativas sobre este conjunto de datos. Estas estadísticas son: valor medio, desviación típica estándar, percentil 25 %, percentil 75 %, valor máximo y valor mínimo. En la figura 3.3 se puede ver la representación gráfica de estos valores.

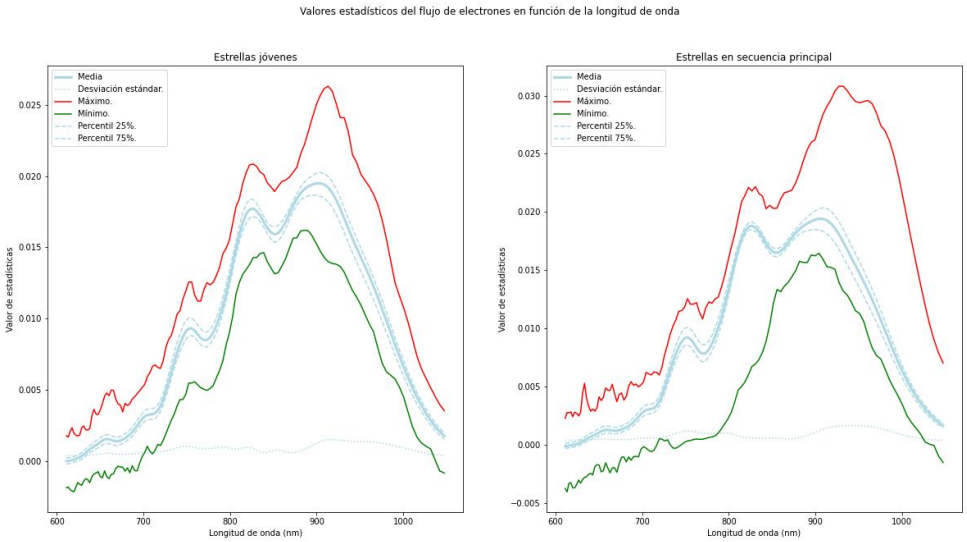

Figura 3.3: Representación de los estadísticos más importantes. Las figuras muestran el valor medio del espectro, la desviación estándar, los valores máximos y mínimos, así como los valores de percentil 25 % y 75 %. La figura de la izquierda las estrellas jóvenes y la figura de la derecha las estrellas en secuencia principal.

Outliers

Se procede a hacer una revisión de valores atípicos en los espectros para las estrellas jóvenes y las estrellas en secuencia principal. Usando una variación del método IQR, en la cual se toma como IQR a la distancia que hay entre el percentil 10 y el percentil 90, se identifican 471 espectros de estrellas en secuencia principal y 42 espectros de estrellas jóvenes que contienen en alguno de sus flujos valores atípicos. La figura 3.4 muestra la representación de estos espectros identificados como atípicos.

Correlaciones

En la figura 3.5 se visualiza en forma de mapa de calor la correlación que existe entre las variables que representa el flujo de electrones producidos en cada CDD por la llegada de luz en cada longitud de onda. Por las propias características estructurarles de los instrumentos, es decir, de los prismas de dispersión y de los CCDs, se ha de esperar una correlación entre los valores de flujo de píxeles vecinos. En el propio proceso físico de dispersión de la luz a través de su paso por los prismas, los fotones que forman la luz se ven desviados de forma continua según la longitud de la onda electromagnética. Por ser una desviación continua, en los bordes de los píxeles vecinos la llegada de fotones

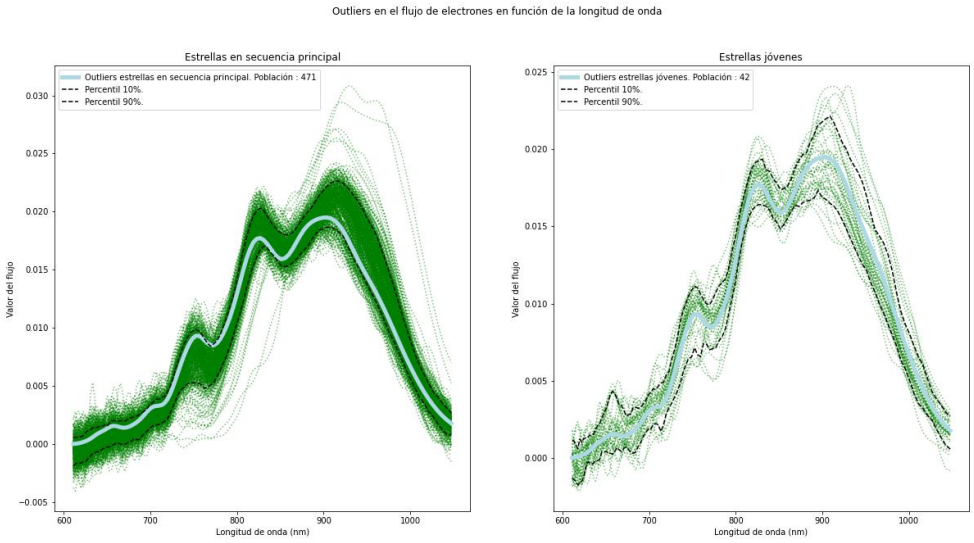

Figura 3.4: Representación de los valores atípicos más importantes. Suele tomarse los percentiles 25 % y 75 % como referencia para la determinación de los valores atípicos. Usando el método IQR, en el que se llama IQR a la distancia que hay entre el percentil 25 % y el percentil 75 %, o lo que es lo mismo, la distancia entre el primer cuartil llamado Q_1 y tercero cuartil llamado Q_2, se determinan como outliers aquellos valores que estén fuera del rango $Q_1 - 1{,}5IQR$, por debajo; y $Q_3 + 1{,}5IQR$, por arriba. Para este conjunto de datos, eso implicaba que más de una tercera parte de las observaciones son atípicas. Se ha reducido los percentiles al 10 % y 90 % por cierto, dejando solo así 471 espectros de estrellas en secuencia principal y 42 de estrellas jóvenes como atípicas. La figura de la izquierda muestra los espectros atípicos de las estrellas en secuencia principal y la figura de la derecha los de las estrellas jóvenes.

sigue una distribución normal, de tal forma que la luz desviada para una longitud de onda será procesada por píxeles adyacentes, provocando así una correlación entre los valores de flujo de cada píxel.

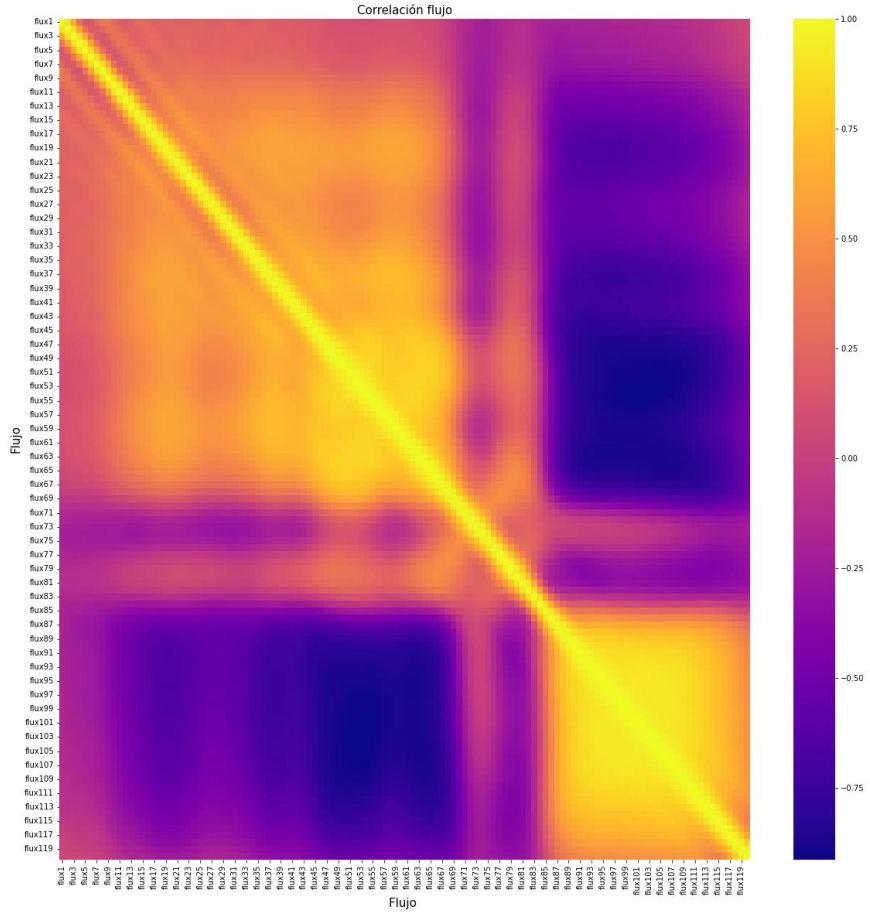

Figura 3.5: Representación de las correlaciones entre los valores de los flujos. Representación con mapa de calor entre los valores de los flujos de espectros recibidos.

Se observan tres regiones principales, una de alta correlación entre los flujos 87 y 120 (zona amarilla situada en la parte inferior derecha), otra también de alta correlación pero menor que la anterior entre los flujos 1 y 70 (zona anaranjada situada en la parte superior izquierda) y, por otro lado, una de mucha menos correlación entre los 70 primeros flujos con los últimos 60 flujos (zona azulada en situada en la parte inferior izquierda).

Capítulo 4

Metodología

Una vez se ha presentado el contexto y objetivo del trabajo, revisado el estado del arte y la procedencia de los datos, se pasa a describir los pasos que se han realizado en el proceso de búsqueda de los espacios latentes. Los pasos que formarán el proceso son:

1. Preprocesado de datos
2. Descripción de las técnicas de obtención de espacios latentes
3. Experimentación y ajustes de parámetros
4. Representaciones gráficas
5. Validación de resultados

4.1. Preprocesado de datos

El lenguaje de programación principal que se ha utilizado para este trabajo ha sido Python 3.9. Básicamente por la facilidad de uso y la gran cantidad de librerías disponibles que existen en la comunidad de desarrollo. La herramienta de desarrollo de código ha sido Jupyter Notebook, igualmente por la facilidad de uso y la gran cantidad de utilidades que ofrece. Una vez establecido el entorno de trabajo correspondiente, la preparación de datos se ha realizado fundamentalmente con la librería Pandas (https://pandas.pydata.org/). Cómo se ha comentado en el apartado de datos, la información con las observaciones está contenida en dos ficheros tipo *.csv*. Uno de ellos con el contenido de las 24.810 estrellas en la secuencia principal y el otro con 3.133 estrellas jóvenes. El preprocesado de datos ha consistido en cinco tareas:

Borrado de registros no válidos

Se cargan en dos DataFrames (estructura fundamental de trabajo de Pandas) el contenido de los dos ficheros. Se hace un limpiado de los registros que contienen información incompleta. Como resultado, en el fichero de estrellas en secuencia principal no se encuentran registros inválidos, pero del fichero de estrellas jóvenes se eliminan 465 registros. Es decir, hay 465 estrellas con los espectros incompletos que no se tendrán en cuenta.

Creación de DataFrame principal

A continuación, se unifican los dos conjuntos de datos en un único DataFrame. El objetivo es que a la hora de llamar a las funciones que se van a encargar de la obtención de los espacios latentes, traten a todas las observaciones a la vez. Así, los nuevos espacios latentes contendrán la información esencial del conjunto de datos formado por todas las observaciones, no de forma separada los creados por las estrellas jóvenes o en secuencia principal.

Clasificación de estrellas

Se procede a crear una nueva variable dentro del DataFrame unificado para diferenciar el tipo de estrellas a la que pertenece cada registro. Esta variable contendrá el valor 1 si el registro viene del fichero de estrellas jóvenes y un 0 si la información proviene del fichero de estrellas en secuencia principal.

Creación DataFrame con medidas Fotométricas

Como el objetivo es encontrar los espacios latentes asociados a las variables fotométricas, se crea un nuevo DataFrame que contiene solo las 120 variables que forman el conjunto de variables fotométricas. Es decir, las relacionadas con los flujos de energía recibida por las cámaras del telescopio del espectro RP.

Cuidado índices

En cada uno de los pasos de manipulación de los DataFrames, se procede a inicializar los índices de los diferentes objetos para no crear duplicados. Es decir, para asegurarse que cada uno de los registros tiene un valor de índice no repetido. Por venir la información en dos ficheros, los índices estarán repetidos para los primeros 2.668 registros, lo que puede dar lugar a error a la hora de manipular los objetos resultantes.

4.2. Descripción de las técnicas de obtención de espacios latentes

Esta sección se centra en describir las técnicas utilizadas para la obtención de los nuevos espacios latentes.

Análisis de Componentes Principales

Probablemente, la técnica de búsqueda de espacios latentes y de análisis de componentes principales más extendida, simple y usada es el llamado PCA. Se puede decir de forma resumida, que está técnica consiste en tomar como base del nuevo espacio latente los autovectores obtenidos de la matriz de covarianzas.

Dado un conjunto inicial de variables (x_1, x_2,...x_p) se busca un nuevo conjunto de variables (v_1, v_2,...v_p) no correladas entre sí y cuyas varianzas vayan decreciendo progresivamente. Así a la hora de elegir las l primeras variables (siendo $l < p$) se preservará la máxima varianza posible.

Cada v_j donde $j \in \{1, 2, ...p\}$, será una combinación lineal de las variables iniciales, es decir:

$$v_j = a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p = \mathbf{a}_j \cdot x \tag{4.1}$$

siendo $\mathbf{a}_j$ un vector de constantes $\mathbf{a}_j = (a_{j1}, a_{j2}, ..., a_{jp})$. Para mantener la ortogonalidad de la transformación de $\mathbf{x}$ en $\mathbf{v}$, se impone que el módulo de dicho vector sea 1. Es decir:

$$\sum_{k=1}^{p} a_{kj}^2 = 1 \tag{4.2}$$

El primer componente se calcula eligiendo aquel a_1 que hace que v_1 tenga la mayor varianza posible, donde la varianza viene dada por:

$$Var(v_1) = Var(\mathbf{a}_1 \cdot x) = \mathbf{a}_1 \cdot \Sigma a_1 \tag{4.3}$$

Un método que se puede usar para maximizar esta función de varias variables sujeta a restricciones, es el método de multiplicadores de Lagrange. En este caso, la restricción es que el módulo sea 1. Se construye la función L de la siguiente forma:

$$L(a_1) = \mathbf{a} \cdot \Sigma a_1 - \lambda(\mathbf{a}_1 \cdot a_1 - 1) \tag{4.4}$$

Maximizando esta función:

$$\frac{\partial L}{\partial \mathbf{a}_1} = 2\Sigma\mathbf{a}_1 - 2\lambda I\mathbf{a}_1 = 0 \Rightarrow (\Sigma - \lambda I)\mathbf{a}_1 = 0 \tag{4.5}$$

Este es un sistema lineal de ecuaciones. Según el teorema de Roché-Frobenius, para que el sistema tenga una solución distinta de 0 la matriz $(\Sigma - \lambda I)$ tiene que ser singular (que no tiene inversa) cuyo determinante es igual a 0. Entonces $|\Sigma - \lambda I\ | = 0$, lo que implica que λ es un autovalor de Σ. Por lo tanto, los autovalores de Σ son aquellos que maximizan la varianza. Se elegirá entonces los autovectores asociados a esos autovalores como la base para formar el nuevo espacio latente, ya que ordenados de mayor a menor serán los que vayan maximizando la varianza en el nuevo espacio. La elección de los a_{p-1} restantes se hace de modo que las variables asociadas v_{p-1} estén totalmente incorreladas con v_1 y entre sí, siguiendo el mismo proceso que para la elección de v_1.

Existe una variación popular de esta técnica que consiste en realizar una Descomposición de Valores Singulares de la matriz de datos X, de tal forma que:

$$\mathbf{X} = \mathbf{U}\ \mathbf{S}\ \mathbf{V}^T \tag{4.6}$$

donde $\mathbf{U}$ es una matriz unitaria, $\mathbf{S}$ es la matriz diagonal de los valores singulares s_i y $\mathbf{V}^T$ será la matriz formada por las direcciones principales del nuevo espacio. Se sabe que la relación entre los valores singulares y los autovectores de la matriz de covarizna viene dada por:

$$\lambda_i = \frac{s_i^2}{(n-1)} \tag{4.7}$$

siendo n el numero de observaciones. Se llamará a esta variedad técnica SVD (por sus siglas en inglés; Singular Value Descomposition) y se usará durante la fase de experimentación.

Autoencoders

Un Autoencoder es una arquitectura especial de red neuronal capaz de aprender las representaciones de los datos de entrada, llamadas representaciones o codificaciones latentes. Estas representaciones tienen normalmente una dimensión menor que la de los datos de entrada, por lo que son especialmente útiles para la reducción de la dimensionalidad y, por lo tanto, para la creación de un nuevo espacio latente.

Esta arquitectura se parece bastante a una arquitectura de un perceptrón multicapa, con la característica principal de que existe una simetría respecto a la capa central. Eso implica que la cantidad de neuronas de la capa de entrada debe de ser igual al número de neuronas de la capa de salida [Gé19]. La forma más simple de un Autoencoder se basa en una red neural de retroalimentación, es decir, la información se mueve en una única dirección (no de forma cíclica) pasando a través de todos los nodos que forman la red, de tal manera que estos se alimentan teniendo en cuenta los pesos y las funciones de activación correspondientes.

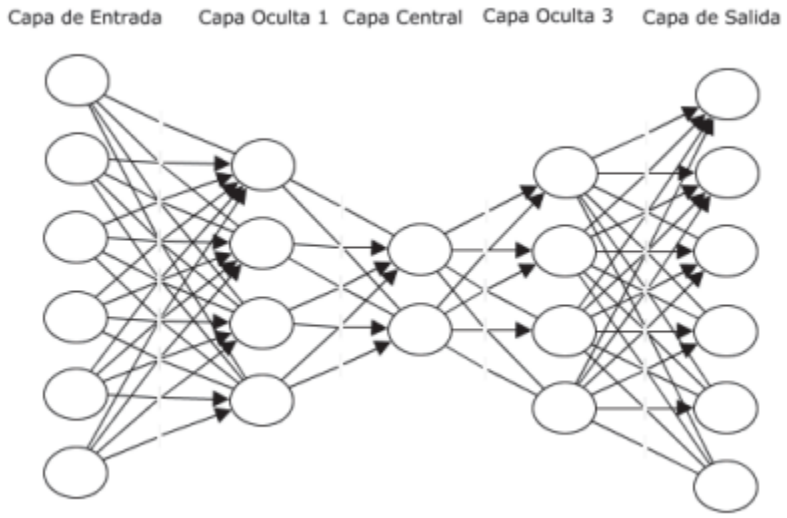

Figura 4.1: Representación esquemática de una red neuronal llamada Autoencoder. Existe una simetría en cuanto al número de capas y de neuronas en cada capa. Este ejemplo consta de dos capas ocultas, dos capas de entrada y salida y la capa central. Las capas adjuntas a la capa de entrada forman el encoder o codificador, las capas adjuntas a la capa de salida forman el decodificador o decoder

Se llama codificador a la mitad de la arquitectura que va desde la capa de entrada hasta la capa central, ambas incluidas, y se llama decodificador a la mitad que va desde la capa central hasta la capa de salida. Será el llamado codificador el elemento que se encargará de aprender las características más importantes de los datos, y la que se usará para crear el nuevo espacio latente. El número de neuronas de la capa central determinará el número de dimensiones que se quiere para el nuevo espacio latente. La esencia de esta arquitectura es que la capa decodificadora aprenda a generar, a partir de los datos codificados, una nueva representación lo más parecida posible a la entrada original. Existen variantes a esta esencia según se quiera obligar a los codificadores a aprender ciertas características útiles de los datos. Algunos ejemplos son los codificadores automáticos regularizados (Sparse , Denoising y Contractive), que son muy utilizados en la codificación de representaciones para tareas posteriores

de clasificación. Otro ejemplo de variante es Autoencoder Variational, muy populares por su utilidad como modelos generativos. Los Autoencoders en general, se utilizan para diversos fines,como por ejemplo,reconocimiento facial o adquisición del significado semántico de las palabras.

Desde el punto de vista formal, si se llama X al espacio inicial y X′ al espacio latente, se puede considerar al codificador y el decodificador como dos transiciones Φ y Ψ respectivamente, de tal forma que:

$$\Phi : X \to X^{'} \tag{4.8}$$

$$\Psi : X^{'} \to X \tag{4.9}$$

En cada capa oculta, la transición de codificación Φ toma la entrada $x \in \mathbb{R}^d = X$ y lo asigna a $x^{'} \in \mathbb{R}^p = X^{'}$, siendo d la dimensión del espacio de entrada y p la dimensión del espacio latente.

La parte de la red codificadora esta formada por una o más capas, para cada paso por cada una de las capas se se tiene la siguiente función de paso :

$$\phi_i = f_{acti}(w_i x + b_i) \tag{4.10}$$

Donde $i \in \{1, 2, ..., N_c\}$, siendo N_c el número de capas del codificador, f_{acti} la función de activación, w_i el peso y b_i el sesgo del salto entre capas. Estos dos último se pueden inicializar de forma aleatoria para luego ir actualizándose durante el proceso de retropropagación.

De la misma forma, la parte decodificadora puede estar compuesta por una o más capas y para cada paso por cada una de las capas se tiene la siguiente función de paso:

$$\psi_i = f'_{acti}(w'_i x + b'_i) \tag{4.11}$$

En este caso, $i \in \{1, 2, ..., N_c\}$ siendo N_c el número de capas del decodificador, f'_{acti}, w'_i y b'_i los nuevos elementos para la red de decodificación.

Como el objetivo es obtener salidas del decodificador que sean lo mas parecidas a las entradas del codificador, se deberá de resolver el siguiente problema de optimización:

$$\Phi, \Psi = \arg\min_{\Phi,\Psi} \|X - (\Phi \circ \Psi)X\|^2 \tag{4.12}$$

Como se mencionó anteriormente, el entrenamiento de un codificador automático se realiza mediante la propagación hacia atrás del error, al igual que una red neuronal de retroalimentación regular. Para entrenar un Autoencoder, se utiliza una función de pérdida que mide la diferencia entre los datos de entrada y los datos de salida. El entrenamiento consiste buscar en cada fase del entrenamiento los parámetros w_i, w'_i, b_i y b'_i que minimicen esta función de pérdida. Se utilizarán algoritmos de optimización, como el descenso del gradiente estocástico. Al final del entrenamiento, el Autoencoder habrá aprendido una representación de menor dimensionalidad que preserve la información relevante de los datos de entrada.

Mapas de Difusión

Dado un conjunto de datos $X_i = x_i$, donde $i \in \{1, 2, ..., N\}$ y cada x_i está compuesto por m variables (es decir, X_i es m-dimensional) se puede construir un grafo (diagrama que representa mediante nodos y arcos las relaciones ente pares de elementos) finito con N nodos de la siguiente forma: cada arco conecta dos nodos cuyos bordes están ponderados a través de un kernel no negativo, simétrico y positivo $w : \mathbf{X} \times \mathbf{X} \to (0, \infty)$.

Por ejemplo, para un kernel típico Gaussiano se define

$$\mathbf{w}(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}) \tag{4.13}$$

donde σ es el parámetro de anchura del núcleo. El núcleo refleja el grado de similitud (grado de ponderación) ente x_i y x_j, y $\| \cdot \|$ es la norma Euclídea en $\mathbb{R}^m$. La matriz resultante $W = w(x_i, x_j)$, de dimensión $N \times N$, se llama matriz de afinidad. Entonces:

$$d(x_i) = \sum_{x_j \in X} w(x_i, x_j) \tag{4.14}$$

se denomina grado de x_i ; se construye la matriz de Markov o de transición llamada $\mathbf{P}$ calculando cada entrada de dicha matriz como:

$$p\left(x_i, x_j\right) = \frac{w\left(x_i, x_j\right)}{d\left(x_i\right)} \tag{4.15}$$

Este término puede interpretarse como la probabilidad de transición de ir del punto x_i al punto x_j . Teniendo en cuenta que se aplicará la función kernel, entonces se puede ver que la probabilidad de transición será alta para elementos cercanos (en este ejemplo, distancias euclídeas pequeñas) y baja para elementos distantes (en este ejemplo, distancias euclídeas grandes). Definiendo $\mathbf{p}^t(x_i, x_j)$ como la probabilidad de transición de ir del punto x_i al punto x_j en un número t de iteraciones o pasos, tenemos entonces la matriz de transición $\mathbf{P}^t$. El cambio de tamaño de t permite controlar la generación de grupos de puntos más específicos o amplios (se controla en número de vecinos a tener en cuenta). Debido a la propiedad de simetría de la función kernel, para cada $t \geq 1$ se pueden calcular un conjunto de N autovalores de $P = \lambda_0 \geq \lambda_1 \cdots \geq \lambda_N$ con sus correspondientes autovectores $\Phi_j, j = 1 \ldots N$ que por definición

$$\mathbf{P}^t \Phi_j = \lambda_j^t \Phi_j \tag{4.16}$$

Usando los autovectores como un nuevo conjunto de coordenadas (que conservan la información del conjunto de datos original) se puede realizar un mapeo desde el conjunto de datos original a un espacio reducido euclídeo l-dimensional (donde l<m) de la siguiente forma:

$$\Psi_t : x_i \to [\lambda_1^t \phi_1(x_i), ...\lambda_l^t \Phi_l(x_i)]^T \tag{4.17}$$

Bastará entonces seleccionar l autovectores para definir el nuevo espacio buscado X′ l-dimensional.

Para este caso en el que se usa el kerenel Gaussiano, no existe una buena teoría para la selección del valor de σ apropiado. Se han propuesto métodos que se reducen

a intercambiar la escasez de la matriz del kernel (pequeño σ) con una caracterización adecuada de la verdadera similitud de dos puntos.

La parte más costosa del mapa de difusión es la construcción de la matriz de afinidad, ya que el calculo se escala cono N^2. Sin embargo, esta matriz es simétrica. Esto significa que solo es necesario calcular $\frac{N^2-N}{2}$ entradas de la matriz. El cálculo de esta matriz podría hacerse en paralelo muy fácilmente, ya que dos entradas cualesquiera son completamente independientes entre sí. La matriz de transición se puede obtener de W dividiendo cada fila por elementos con $\mathrm{d}(x_i)$. Esto también podría hacerse en paralelo. Los resultados experimentales muestran que la matriz resultante es dispersa, y dado que solo se necesita encontrar los primeros autovectores propios y autovalores, esto no representa un gran problema de computación.

4.3. Experimentación y ajustes de parámetros

En esta sección se describe el proceso de experimentación con las opciones de ejecución de las tres técnicas descritas, con el objetivo de encontrar aquellas que mejor resultado produzcan. Es bien conocido que en la implementación de herramientas de Machine Learnig existe una fase importante que se centra en la búsqueda de aquellos hiperparámetros de los modelos que optimicen los resultados obtenidos y, por lo tanto, determinen las mejoras maneras de usar dichas herramientas. El caso que ocupa este trabajo puede entenderse que de forma similar, debe de implementarse una fase en la que se buscan aquellas opciones de ejecución de las técnicas descritas que optimicen los resultados.

Análisis de Componente Principales

Por definición, el Análisis de Componentes Principales no tiene parámetros. Es decir, es una técnica no paramétrica. No obstante, como se indicaba en el apartado anterior, se experimenta con la variedad SVD que proporciona la librería sklearn.decomposition.PCA [slhls] que se ha utilizado para este trabajo. Esta posee cierta configuración opcional que puede usarse para buscar diferentes soluciones. Las opciones sobre las que se ha experimentado son:

- **Whiten**. Opciones disponibles *True* y *False*. Cuando está activada esta opción (*True*) los autovectores resultantes son multiplicados por la raíz cuadrada del número de observaciones y divididos por los valores singulares. Esto elimina parte de la información de los datos (las escalas de varianza relativa de los componentes), pero puede mejorar la precisión predictiva de estimadores posteriores.

- **SVD solver**. Opciones disponibles *full* , *arpack* , *randomized* y *auto*.

  randomized: Se selecciona SVD aleatoria. Este algoritmo encuentra una descomposición aproximada de valores singulares truncados (normalmente muy buena) utilizando la aleatorización para acelerar los cálculos. Es particularmente rápido en matrices grandes en las que desea extraer solo una pequeña cantidad de componentes.

aut: El solver se selecciona mediante una política predeterminada basada en el tamaño de los datos y las variables de entrada. Si el tamaño de los datos de entrada tiene más de 500 observaciones con más de 500 variables y la cantidad de componentes para extraer es inferior al 80 % de la dimensión más pequeña de los datos, entonces se habilita el método "randomized". De lo contrario, se calcula el SVD completo exacto.

full: El solver es SVD llamando al solucionador LAPACK estándar scipy.linalg.svdy y luego selecciona los componentes mediante postprocesamiento.

arpack: El solver es SVD truncando en el número de componentes elegido llamando al solver ARPACK a través de scipy.sparse.linalg.svds. Requiere estrictamente que el número de componentes sea mayor que cero y menor que el número de observaciones y menor que el número de variables originales

Existen otras opciones muy específicas para cada uno de los solvers comentados, pero no se han tenido en cuenta para esta fase de experimentación, ya que como se comentaba, esta técnica no es paramétrica y la finalidad de usarla es para tenerla como punto de comparación con las técnicas no lineales.

Autoencoders

Una característica fundamental de las arquitecturas de las redes neuronales es la cantidad de capas ocultas que pueden llegar a tener. Por ese motivo, en este trabajo se han probado diferentes arquitecturas con diferentes números de capas en cada parte que compone a un Autoencoder.

Como se comentaba en el apartado anterior, en la fase de entrenamiento se busca aquellos parámetros de la función de activación que minimizan la función de perdida definida. Para la búsqueda de dichos mínimos, se puede contar con diferentes algoritmos, y son estos con los que se han experimentado en este trabajo. La librería utilizada para la creación de esta red ha sido Keras que proporciona tensorflow. Existe una variedad de algoritmos de optimización dispuestos por esta librería, como por ejemplo: SGD, RMSprop, Adam, AdamW, Adadelta, Adagrad, Adamax, Adafactor, Nadam y ftrl. Por ser los más usados por sus resultados en varias materias, se han tomado Adam, RMSprop y SGD.

Por último, intrínsecamente relacionado con el algoritmo de optimización, se prueban varios valores de la tasa de aprendizaje (Learning Rate) que varían entre 10^{-4} y 1. El learning rate establece cómo de rápido pueden variar los parámetros que minimizarán la función de perdida a medida que se va entrenando la red. Este hiperparámetro es uno de los más difíciles de fijar, ya que tiene una dependencia muy grande con los datos y puede afectar enormemente al resto de hiperparámetros de la red. Si el learning rate es muy grande, el proceso de optimización puede ir saltando de una región a otra sin detenerse en los mínimos que puede haber entre dichas regiones. Por el contrario, si es muy pequeño, el proceso de entrenamiento puede requerir de muchísimas iteraciones para encontrar el mínimo e incluso, no llegar a completarse.

Mapas de Difusión

Atendiendo a la definición teórica que se ha hecho en el apartado anterior, sobre la técnica Mapas de Difusión, se identifican tres opciones que se podrán ajustar y experimentar a la hora de obtener los nuevos espacios latentes:

- σ: En el caso de usar un núcleo Gaussiano, como se ha puesto de ejemplo en el apartado anterior, existe el parámetro σ el cual actúa sobre el núcleo de forma que cuando más grande es, más peso se asigna a las observaciones próximas. En este trabajo se usa exclusivamente el núcleo Gaussiano y se ha experimentado con los siguientes valores de σ : 0.5, 1 y 2

- α: Algunas implementaciones de esta técnica modifica ligeramente la definición de la matriz de transición (4.15) elevando el denominador por α para ajustar la influencia de la densidad de puntos en las transiciones. Si no se sabe si el muestreo de las observaciones está relacionado con el nuevo espacio, se establece $\alpha = 1$ (el operador de difusión se aproxima al operador de Laplace-Beltrami). Si se pretende recuperar la geometría riemanniana del conjunto de datos, independientemente de la distribución de los puntos, puede usar $\alpha = 0.5$ (la cadena de Markov resultante se aproxima a la difusión de Fokker-Planck). Y con $\alpha = 0$ se reduce a la normalización laplaciana gráfica clásica. Estos son los valores de α con los que se ha experimentado.

- **t**: Este parámetro especifica el número de transiciones entre puntos a la hora de crear la matriz de probabilidad, controlando la generación de grupos más específicos o amplios. Se han tomado los valores : 1 y 1000.

4.4. Representaciones gráficas

Antes de proceder a visualizar los puntos (observaciones) en los nuevos espacios latentes para cada una de las tres técnicas descritas, se procede a añadir a los DataFrames resultantes dos variables. A cada una de las observaciones se le añaden la temperatura y la clase de estrella, es decir 0 o 1 según sea estrella en secuencia principal o joven. El objetivo de añadir esta información a los DataFrames obtenidos es poder analizar visualmente los puntos en los nuevos espacios latentes, distinguiendo mediante símbolos el tipo de estrella y mediante escala de colores la temperatura. De forma que, reduciendo la dimensión del espectro fotométrico de 120 variables a 9 variables, se puedan buscar aquellas nuevas variables que, enfrentadas en una gráfica de dispersión, permitan diferenciar visualmente las estrellas jóvenes de las no jóvenes. Si se consigue de alguna forma que, para alguna de las nuevas bases de los espacios latentes, los puntos tengan valores diferentes para las estrellas jóvenes, de las no jóvenes, serán esas nuevas bases las que estén recogiendo o indicando de alguna forma la edad de la estrella.

Las representaciones gráficas que se realizarán serán las siguientes:

1. Diferencia de medianas
2. Boxplot
3. Diagrama de dispersión
4. Densidad de población

4.4.1. Diferencia de medianas

Como resultado de la fase de experimentación, se tiene un conjunto de s_i nuevos espacios con nuevas variables para cada una de las técnicas. Para poder discernir cuáles de dichas técnicas están produciendo los mejores resultados, se representa gráficamente la diferencia entre las medianas de cada tipo de estrella para cada una de las nuevas variables. Las diferencias de las medianas están indicando cuanto de alejados son las nuevas variables para los dos tipos de estrellas. Cuanto más grande son las diferencias, mejor es la técnica para la búsqueda de índicos de juventud en los espectros.

4.4.2. Boxplot

Una vez que se ha identificado para cada técnica cuál es la combinación de parámetros que mejores resultados produce, se procede a representar mediante Boxplot (también llamado, Diagrama de Cajas o de Bigotes) las nuevas variables. El objetivo de dicha representación consiste en ser un poco más riguroso en cuanto a como de separadas están las nuevas variables para cada tipo de estrella, no fijándose únicamente en la mediana, sino también observar como se comportan otras variables estadísticas como los percentiles o los outliers. También sirve esta representación para identificar las dos nuevas variables que mejor recogen los indicios de juventud en los espectros.

4.4.3. Diagrama de dispersión

Se representa en un diagrama de dispersión las dos variables que se han identificado como las que mejor recogen los indicios de juventud de los espectros. Se añade la temperatura como una dimensión más, para que a través de una escala de color, se pueda apreciar donde se sitúan (en el plan formado por estas nuevas variables) las estrellas según su tipo y temperatura.

4.4.4. Densidad de población

Por último, en una representación parecida a la anterior, se muestra la densidad de probabilidad de ocurrencia de valores de las observaciones en las nuevas variables. El objetivo es visualizar donde se sitúan los valores más probables de las dos nuevas variables para cada tipo de estrella.

4.5. Validación de resultados

A continuación se describe a grandes rasgos las principales tareas que se han realizado para comprobar y probar, que los resultados obtenidos son coherentes y que, por lo menos, no se han producido grandes fallos a la hora de realizar el proceso de búsqueda de los espacios latentes:

1. En las diferentes representaciones gráficas se han identificado las estrellas más frías y más calientes de cada tipo de estrella para comprobar que efectivamente los colores de las representaciones eran correctos y no se estaba produciendo ningún error a la hora de hacer las gráficas. Es decir, que cada punto representado en los nuevos espacios corresponden realmente con cada punto del espacio original. Se puede entender esta prueba como una prueba unitaria que identifica los extremos de las observaciones en los nuevos espacios.

2. En las diferentes representaciones gráficas se han identificado los valores máximos y mínimos obtenidos para el conjunto de puntos, para cada una de las nuevas dimensiones, para comprobar que no se están produciendo errores a la hora de realizar las representaciones gráficas. Se puede entender esta prueba como unitaria de la representación gráfica.

3. En relación con el proceso de análisis de los datos y descubrimiento de las principales estadísticas, se han comprobado los resultados obtenidos comparándolos con los obtenidos con otras herramientas (con R y Excel). El objetivo principal, aparte de la propia comprobación de las estadísticas, era comprobar que no se estaba cometiendo error a la hora de que Python interpretara los tipos de datos, como por ejemplo, suele pasar con los separadores de decimales o miles.

4. Se han realizado pruebas unitarias sobre las librerías que se han usado para hacer los mapas de difusión. Es decir, se ha tomado conjunto de datos sencillos para un espacio con un número de dimensiones pequeño y se ha seguido la teoría indicada en el apartado de metodología para comprobar que las librerías dan como resultado el mismo que el modelo teórico.

5. Se han hecho pruebas sobre subconjuntos de datos contenidos en el conjunto proporcionado para verificar que las librerías que se han usado para hacer los mapas de difusión daban resultados consistentes. Es decir, se ha tomado conjunto de datos aleatorios contenidos en el original, se ha procedido a realizar las búsqueda de los espacios latentes y se ha validado que los resultados eran consistentes entre las ejecuciones.

6. Por último, se ha contado con un conjunto extra de 12 ficheros con los espectros de estrellas pertenecientes a diferentes zonas de formación estelar. Algunas de estas se sabe que son zonas de formación estelar jóvenes y otras son viejas y no deberían diferenciarse de la secuencia principal. Se pretende, con este nuevo conjunto de datos, localizar las estrellas de estas zonas en el espacio latente producido por Mapas de Difusión para corroborar que las estrellas pertenecientes a zonas de formación estelar se encuentran localizadas en las zonas esperadas.

Capítulo 5

Resultados

En este capítulo se presentan los resultados obtenidos en la la fase de experimentación y las representaciones gráficas, descritas en el apartado anterior, para cada una de las técnicas.

5.1. Análisis de componentes principales

En la tabla 5.1 se muestran las combinaciones utilizadas para la búsqueda de las mejores soluciones cuando se obtienen los componentes principales con PCA SVD. En total se ha obtenido 8 nuevos espacios, con sus correspondientes 9 componentes principales. Se llamará $PCAn$ a cada una de las nuevas componentes obtenidas.

Solver	**withen** *true*	**withen** *false*
auto	True_auto	False_auto
full	True_full	False_full
arpack	True_arpack	False_arpack
randomized	True_randomized	False_randomized

Cuadro 5.1: Nombre de las combinaciones utilizadas para diferenciar entre la opción seleccionada de solver y withen. Se concatena _ con el valor de *withen* y el valor del *solver*. Así, por ejemplo, la combinación *withen* = *True* y *solver* = *randomized* se llamará True_randomized

La figura 5.1 muestra las diferencias entre la mediana de las estrellas jóvenes y la mediana de las estrellas en secuencia principal, para cada una de las nuevas variables. Se diferencian los resultados por rangos de temperatura. En total se han establecido 8 rangos de temperaturas para agrupar las estrellas. Todos los rangos abarcan 100 K, comenzando por 2000 K y terminando en 2700 K, a excepción del primer rango que abarca temperaturas inferiores a los 2000 K. Hay que comentar que la estrella joven más fría que se tiene es de 2155 K, por lo que los dos primeros rangos no contienen estrellas jóvenes.

La combinación que mejores resultados ofrece, es decir, que muestra mayores diferencias de medianas, es la que combina whiten igual *True* con el método *randomized*. Se observa que las componentes $PCA2$, $PCA3$,$PCA4$, $PCA5$ y $PCA8$ son las que tienen diferencias mayores que 0. Siendo $PCA3$ y $PCA4$ las que más diferencia presentan. El resto parece que no consiguen recoger los indicios de juventud de los espectros.

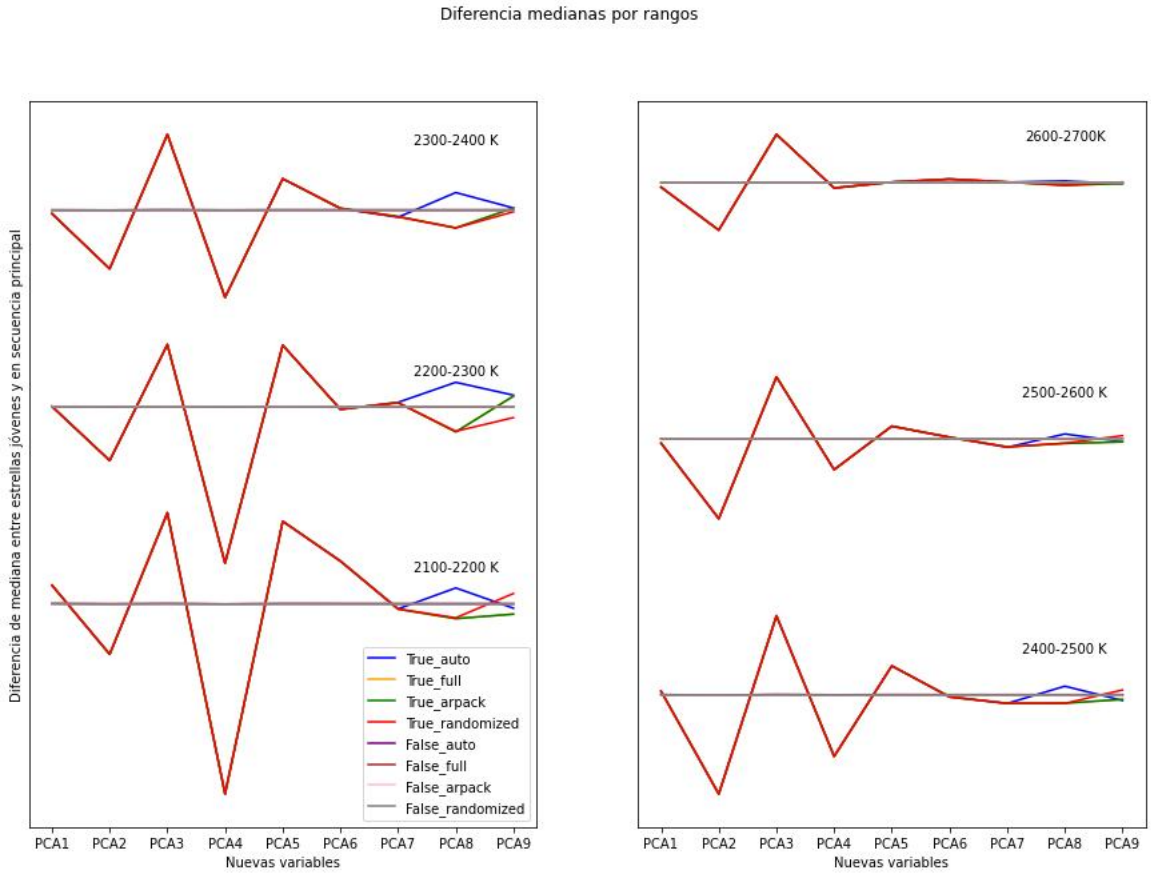

Figura 5.1: Representación gráfica de la diferencia de las medianas entre estrellas jóvenes y en secuencia principal en cada una de las nuevas variables obtenidas con PCA. Las diferencias entre las medianas se muestran por rangos de temperatura que abarcan 100 K comenzando en los 2000 K y terminando en 2700 K (a excepción del primer rango que abarca temperaturas inferiores a 2000 K). Se observa que la combinación *whiten* igual a $True$ y solver igual a *randomized* es la que ofrece la solución más óptima (mayor diferencia de medianas). Las nuevas variables $PCA2$, $PCA3$, $PCA4$, $PCA5$ y $PCA8$ tienen diferencias diferentes de 0, siendo máximas para $PCA3$ y $PCA4$. Cuanto mayor es la temperatura de las estrellas, menor diferencia aparece en las medianas, por lo que es más difícil captar los indicios de juventud de los espectros.

También se observa que cuanto más frías son las estrellas, más grandes son las diferencias. Cuando las temperaturas de las estrellas están en el rango de 2600 - 2700 K las diferencias de las medianas son mucho menores.

Una vez seleccionada la combinación más óptima, se procede a hacer una visualización más rigurosa con boxplot. En la figura 5.2 se muestran las nuevas componentes con detalle de cada uno de los rangos de temperaturas. Las estrellas jóvenes se muestran en rojo y las estrellas en secuencia principal en azul. Se observa que las nuevas variables $PCA3$ y $PCA4$ separan mejor los tipos de estrellas, serán estas sobre las que se profundizará en las figuras siguientes.

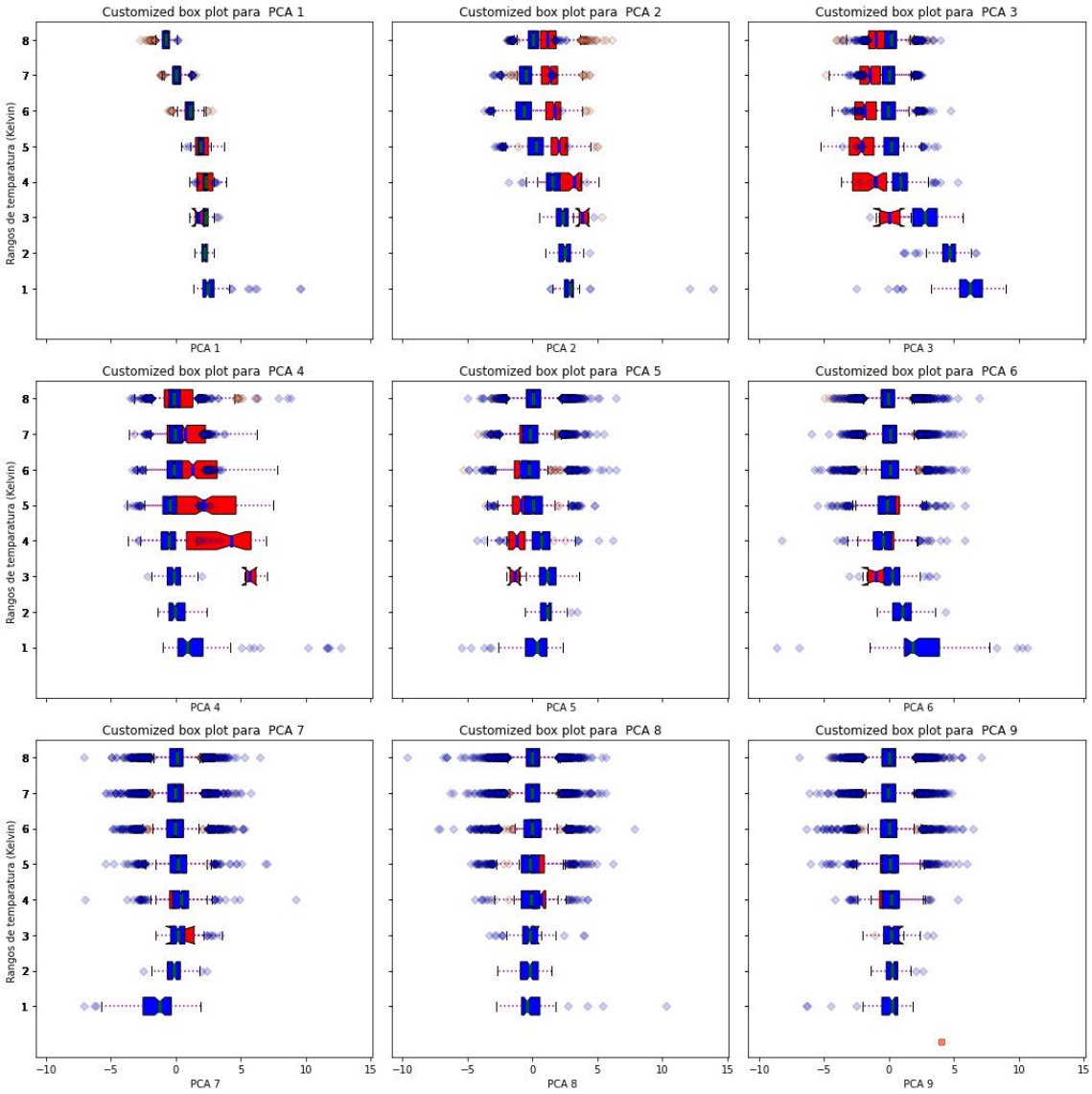

Figura 5.2: BoxPlot por rangos de temperatura de los valores de las observaciones en cada una de las nuevas variables obtenidas con PCA. Se observa que hay mayor separación de valores entre estrellas jóvenes (valores en rojo) y estrellas en secuencia principal (valores en azul) para las componentes $PCA2$,$PCA3$ y $PCA4$.

En la figura 5.3 se muestran los gráficos de dispersión de las observaciones en las componentes principales que mejor han recogido los indicios de juventud. En general existe un grupo importante de estrellas en secuencia principal que tienen valores diferentes de las estrellas jóvenes.

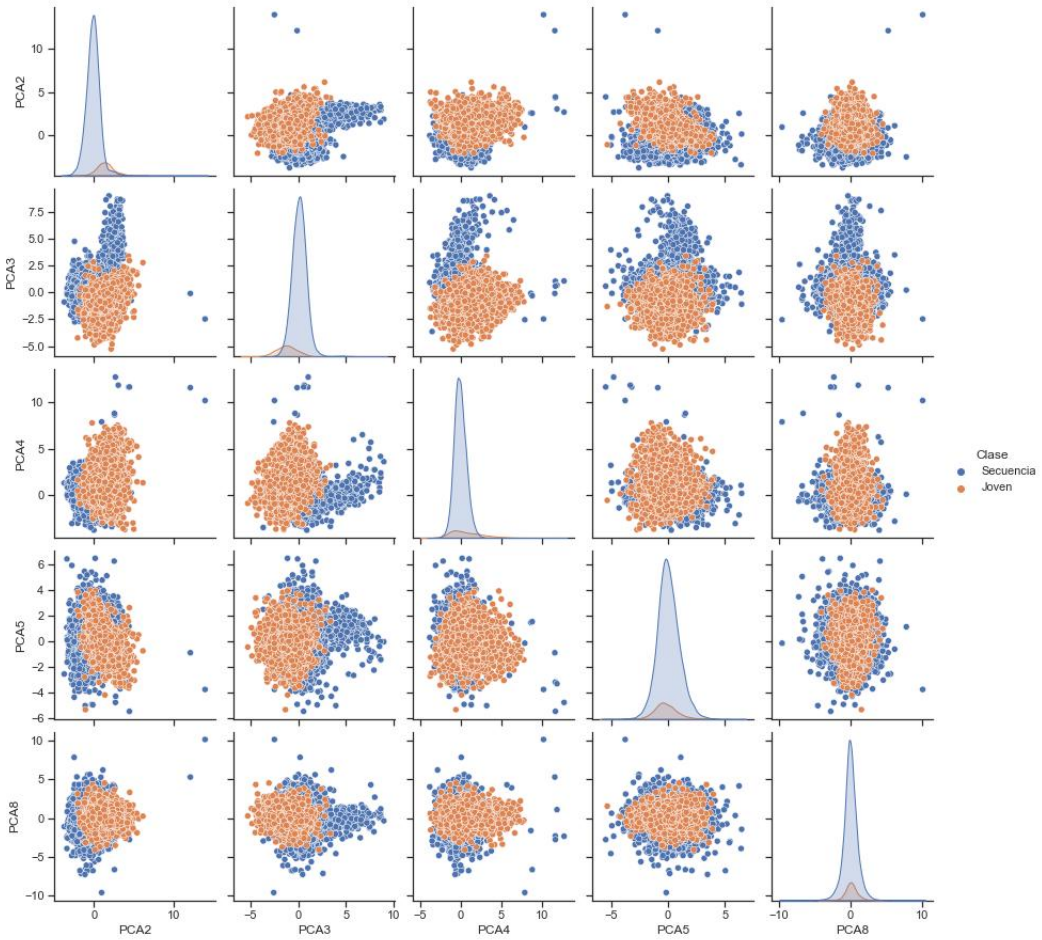

Figura 5.3: Representación en dos dimensiones de las componentes principales extraídas con PCA. En naranja se muestran las estrellas jóvenes y en azul las estrellas en secuencia principal. Existe un conjunto de estrellas jóvenes que muestran valores diferentes en alguna componente a los valores de las estrellas en secuencia principal. Pero en general, casi todas las estrellas (jóvenes y en secuencia principal) se localizan en una misma región. Para estas estrellas, las componentes principales obtenidas con PCA no están recogiendo los indicios de juventud de sus espectros. Parece que $PCA4$ es la componente que mejor recoge los indicios en juventud.

En la figura 5.4 se muestra la representación gráfica de $PCA4$ frente a $PCA3$. Se ha añadido la temperatura de la estrella para poder visualizar como se distribuyen en el nuevo espacio las estrellas según la temperatura. Las estrellas jóvenes tienen valores altos de $PCA4$ y las estrellas en secuencia principal frías tienen valores altos de $PCA3$.

Figura 5.4: Representación gráfica de las características principales $PCA4$ vs. $PCA3$ obtenidas con PCA. Existe una zona donde se acumulan una gran mayoría de puntos y donde no se puede distinguir el tipo de estrella. Se observa que los valores más grandes de $PCA3$ están reservados para estrellas en secuencia principal más frías y que los valores más grandes de $PCA4$ están reservados (en general) para las estrellas jóvenes. Existe un grupo de 6 estrellas en secuencia principal muy frías que se sitúan en los valores altos de $PCA4$ reservado para las estrellas jóvenes.

En la figura 5.5 se representa la densidad de probabilidad de ocurrencia de los valores $PCA4$ y $PCA3$ para los dos tipos de estrellas. Se puede observar como las zonas de alta densidad están muy cercanas, lo que produce un solape para un un conjunto considerable de estrellas. Esto ocurre cuando la temperatura de las estrellas es mas alta.

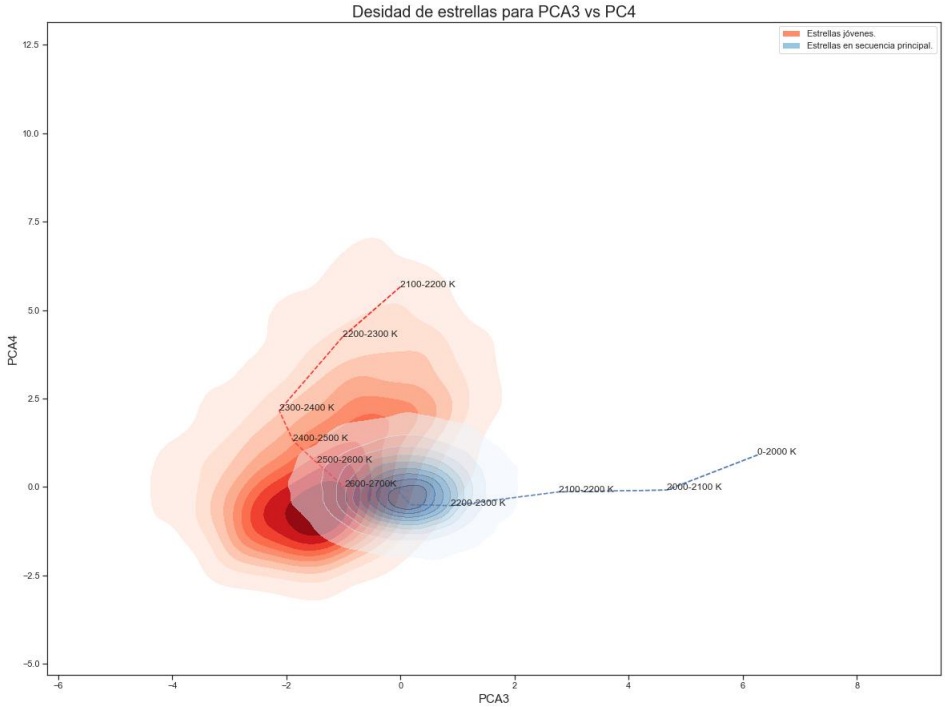

Figura 5.5: Representación gráfica de la distribución de densidad de probabilidad para los valores de $PCA4$ y $PCA3$ de las estrellas jóvenes y en secuencia principal. Esta densidad viene representada como una zona sombreada en color rojizo para las estrellas jóvenes y azulado para las estrellas en secuencia principal. Se puede observar que las zonas de alta densidad (zona saturada de colores) están muy próximas. No se muestra los valores de las densidades de probabilidad, ya que lo interesante no es en sí los valores, sino identificar las zonas de alta densidad que están ocupando los dos tipos de estrellas. Para estimar la densidad se ha usado Kernel Density Estimation que es un método no paramétrico que estima la función de densidad de variables aleatorias a partir de un número finito de observaciones, siendo esta función de densidad continua. En este caso la estimación se ha apoyado en un kernel con base Gaussiana.

5.2. Autoencoders

A continuación se muestran los resultados obtenidos con Autoencoders. Se han utilizado dos arquitecturas diferentes. La primera consiste en una arquitectura compuesta de 4 capas: la primera capa crea un espacio nuevo de 100 variables, la segunda baja hasta las 50 variables, la tercera hasta 25 y la última a las 9 buscadas. La segunda arquitectura solo consta de una capa, es decir, directamente se construye el espacio de 9 variables desde las 120 variables iniciales.

5.2.1. Arquitectura compuesta

En la tabla 5.2 se muestran las combinaciones utilizadas para la búsqueda de las mejores soluciones creadas con Autoencoders. En total se ha obtenido 15 nuevos espacios, con sus correspondientes 9 nuevas variables. Se llamará AUn a cada una de las nuevas variables obtenidas.

	Learning Rate				
Solver	**0.0001**	**0.001**	**0.01**	**0.1**	**1**
SGD	SGD0.0001	SGD0.001	SGD0.01	SGD0.1	SGD1
RMSprop	RMSprop0.0001	RMSprop0.001	RMSprop0.01	RMSprop0.1	RMSprop1
Adam	Adam0.0001	Adam0.001	Adam0.01	Adam0.1	Adam1

Cuadro 5.2: Nombre de las combinaciones utilizadas para diferenciar las opciones de ejecución de las arquitecturas de Autoencoders. Se concatena el valor del *solver* y el valor del *learning rate*. Así, por ejemplo, la combinación de *RMSprop* y un learning rate= 1 se llamará RMSprop1.

La figura 5.6 muestra la diferencias entre las medianas de los valores tomados por las observaciones en cada una de las nuevas variables.

La combinación que mejores resultados ofrece, es decir, que muestra mayores diferencias de medianas, es cuando se usa el optimizador SGD con un learning rate de 0.01. Se observa que las variables $AU1$, $AU2$, $AU4$, $AU6$ y $AU7$ presentan las máximas diferencias. Cuanto más fría es la estrella, más grandes son las diferencias. De nuevo, cuando las temperaturas de las estrellas están en el rango de 2600 - 2700 K las diferencias de las medianas son mucho menores.

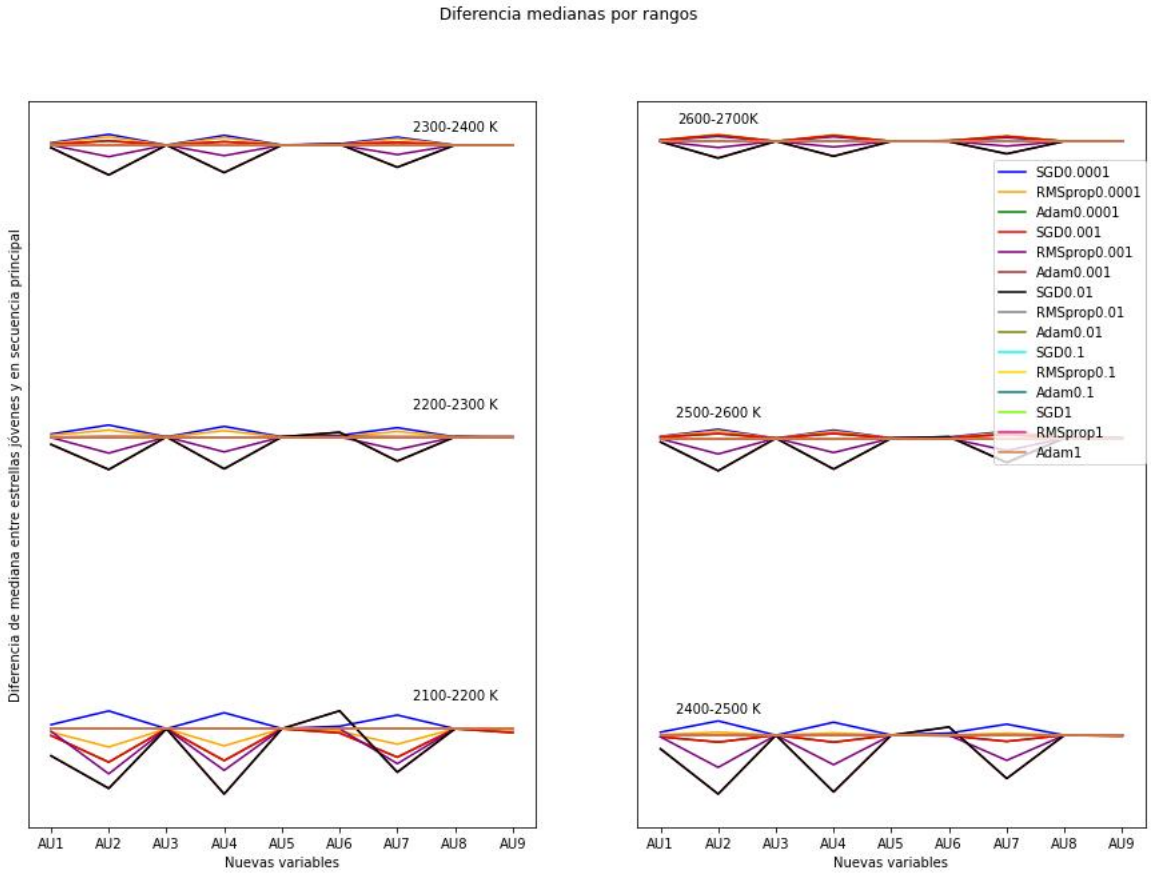

Figura 5.6: Representación gráfica de la diferencia de las medianas entre estrellas jóvenes y en secuencia principal en cada una de las nuevas variables obtenidas con Autoencoders de arquitectura compuesta. Las diferencias entre las medianas se muestran por rangos de temperatura que abarcan 100 K comenzando en los 2000 K y terminando en 2700 K (a excepción del primer rango que abarca temperaturas inferiores a 2000 K). Se observa que la combinación de un learning rate de 0.01 con el optimizador SGD es la que ofrece la solución más óptima (mayor diferencia de medianas). Las nuevas variables $AU1$, $AU2$, $AU4$, $AU6$ y $AU7$ tienen diferencias diferentes de 0, siendo máximas para $AU4$ y $AU2$. Cuanto mayor es la temperatura de las estrellas, menor diferencia aparece en las medianas, por lo que es más difícil captar los indicios de juventud de los espectros.

Una vez seleccionada la recombinación mas óptima se procede a ha hacer una visualización mas rigurosa con boxplot. En la figura 5.7 se muestran las nuevas componentes con detalle de cada uno de las rangos de temperaturas.

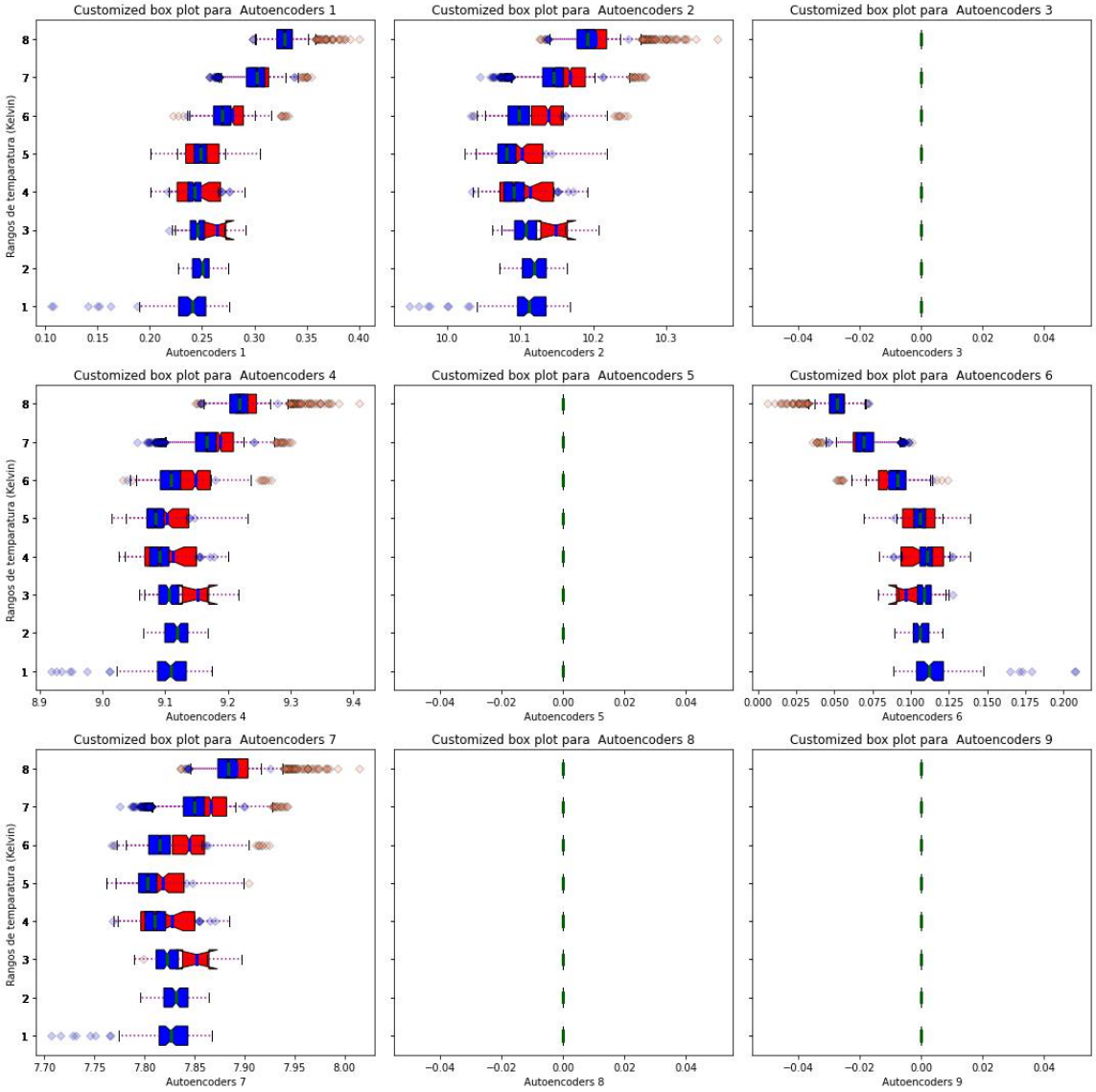

Figura 5.7: BoxPlot por rangos de temperatura de los valores de las observaciones en cada una de las nuevas variables obtenidas con Autoencoders de arquitectura compuesta. Se observa que hay mayor separación de valores entre estrellas jóvenes (valores en rojo) y estrellas en secuencia principal (valores en azul) para las componentes $AU1$, $AU2$, $AU4$, $AU6$ y $AU7$.

En la figura 5.3 se muestran los gráficos de dispersión de las variables que mayor diferencia de medianas han tenido.

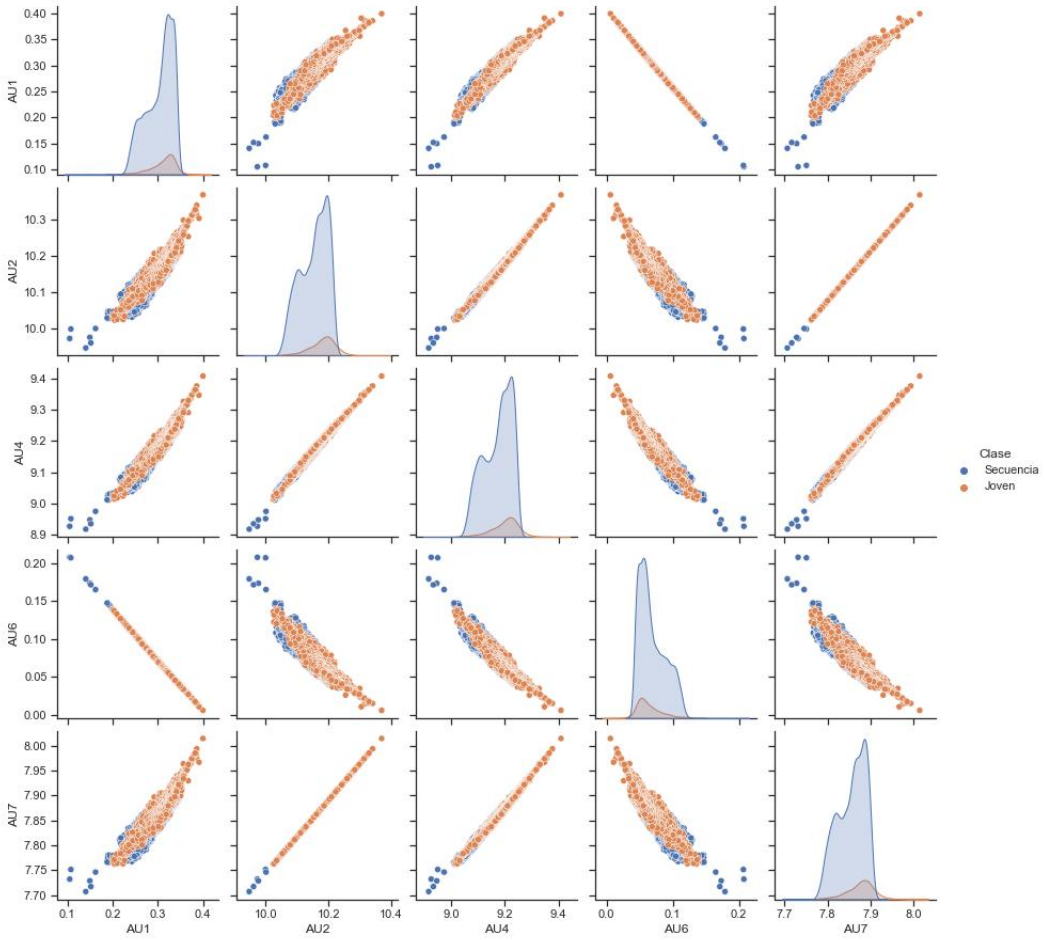

Figura 5.8: Representación en dos dimensiones de las observaciones en el nuevo espacio latente creado con Autoencoders de arquitectura compuesta. En naranja se muestran las estrellas jóvenes y en azul las estrellas en secuencia principal. En general, casi todas las estrellas (jóvenes y en secuencia principal) se localizan en una misma región. El nuevo espacio no están recogiendo claramente los indicios de juventud de sus espectros.

En la figura 5.9 se muestra la representación gráfica de $AU6$ frente a $AU2$. Parece que estas dos nuevas variables son las que mejor separan en el nuevo espacio los espectros de las estrellas en secuencia principal de las jóvenes. No obstante, existe una zona con una alta acumulación de puntos, de donde no se puede distinguir el tipo de estrella. Parece que hay cierta tendencia de localización de estrellas frías en secuencia principal en valores altos de $AU6$ y de $AU2$ para las estrellas calientes jóvenes.

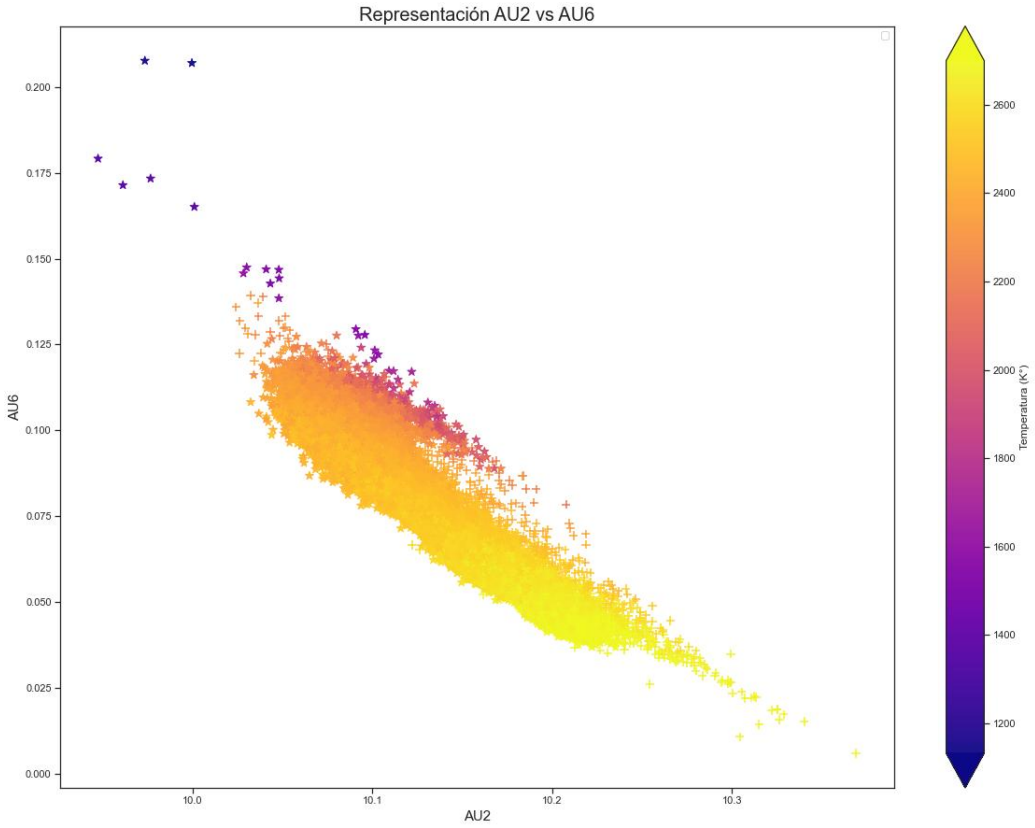

Figura 5.9: Representación gráfica de las nuevas variables $AU7$ vs. $AU2$ obtenidas con Autoencoders. Existe una zona donde se acumulan una gran mayoría de puntos y donde no se puede distinguir el tipo de estrella. Se observa que los valores más pequeños de $AU2$ están reservados para estrellas en secuencia principal fías y que los valores más grandes para algunas de las estrellas jóvenes. Existe un grupo de 6 estrellas en secuencia principal muy frías que se sitúan en los valores pequeños de $AU2$.

En la figura 5.10 se representa la densidad de probabilidad de ocurrencia de los valores $AU6$ y $AU2$ para los dos tipos de estrellas. Se puede observar como las zonas de alta densidad están muy cercanas, lo que produce un solape para un rango considerable de valores. Parece que hay cierta tendencia de acumulación de estrellas jóvenes en valores altos de $AU6$ y $AU2$ e igualmente de estrellas en secuencia principal para valores pequeños de $AU6$ y $AU2$. La imagen no muestra claramente donde se sitúa la densidad de probabilidad de ambos tipos de estrellas.

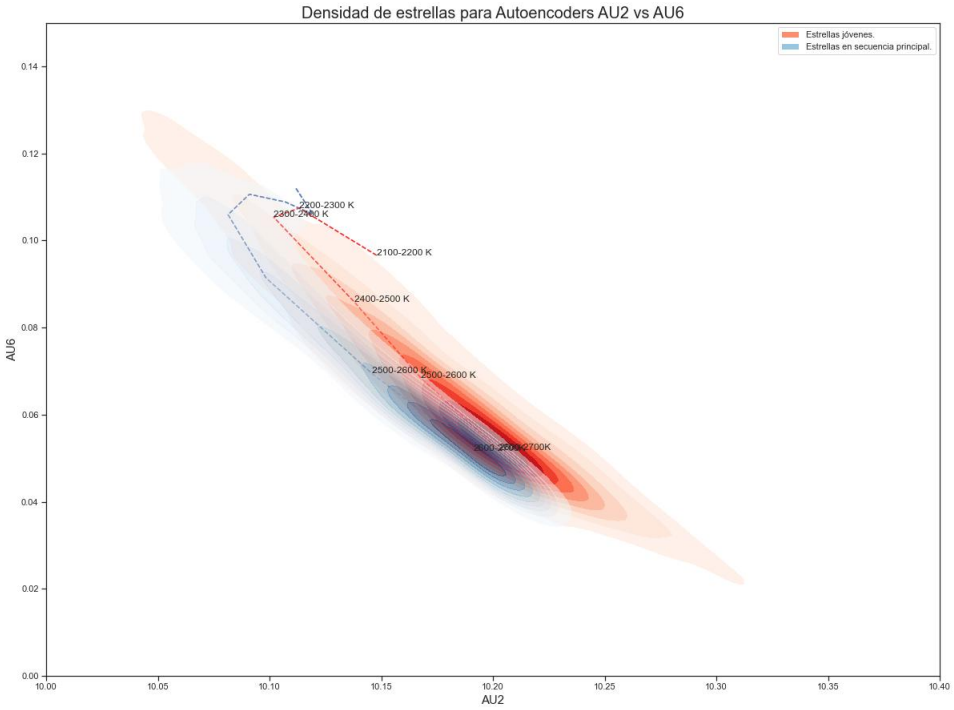

Figura 5.10: Representación gráfica de la distribución de densidad de probabilidad para los valores de $AU6$ y $AU2$ de las estrellas jóvenes y en secuencia principal. Esta densidad viene representada como una zona sombreada en color rojizo para las estrellas jóvenes y azulado para las estrellas en secuencia principal. Se puede observar que las zonas de alta densidad (zona saturada de colores) están muy próximas. No se muestra los valores de las densidades de probabilidad, ya que lo interesante no es en sí los valores, sino identificar las zonas de alta densidad que están ocupando los dos tipos de estrellas. Para estimar la densidad se ha usado Kernel Density Estimation que es un método no paramétrico que estima la función de densidad de variables aleatorias a partir de un número finito de observaciones, siendo esta función de densidad continua. En este caso la estimación se ha apoyado en un kernel con base Gaussiana.

5.2.2. Arquitectura simple

En la tabla 5.2 también se muestran las combinaciones utilizadas para la búsqueda de las mejores soluciones creadas con Autoencoders de una arquitectura simple. En total se ha obtenido 15 nuevos espacios, con sus correspondientes 9 nuevas variables. La figura 5.11 muestra la diferencias entre las medianas de los valores tomados por las observaciones en cada una de las nuevas variables.

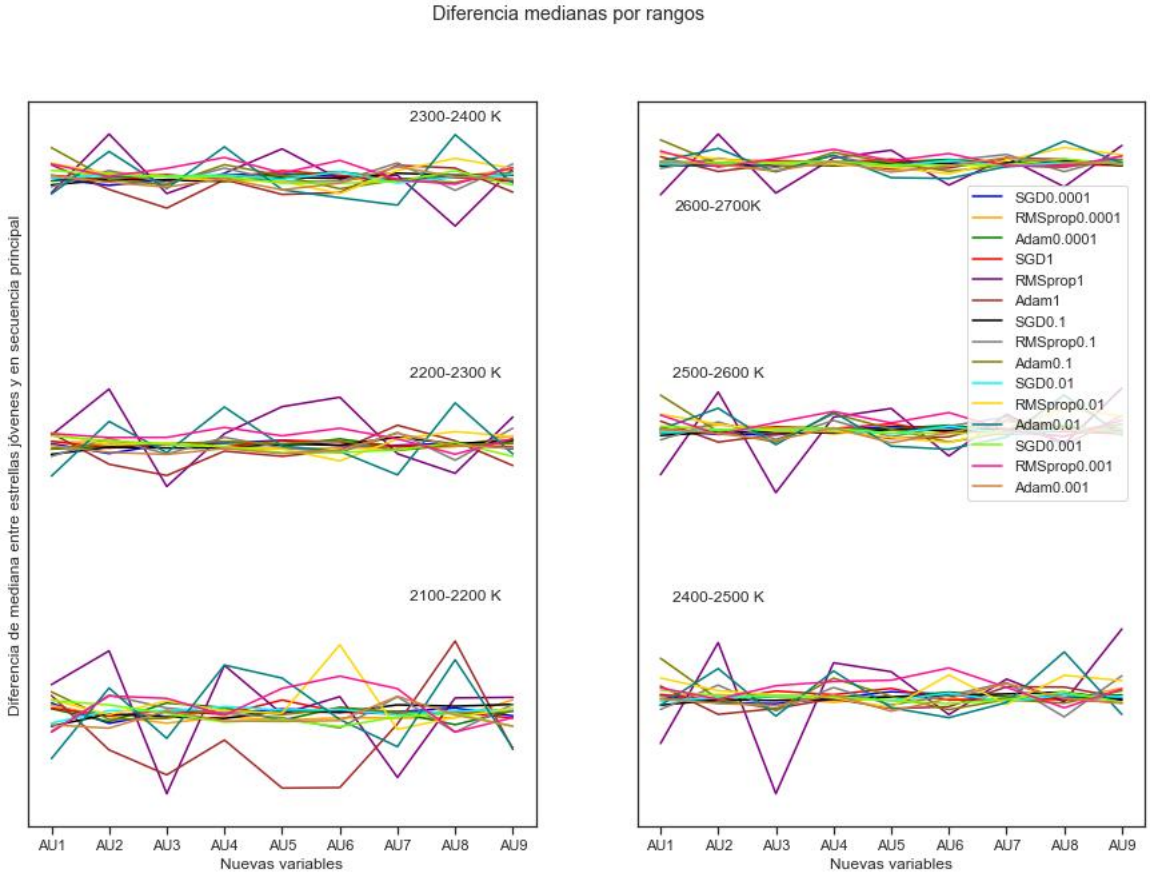

Figura 5.11: Representación gráfica de la diferencia de las medianas entre estrellas jóvenes y en secuencia principal en cada una de las nuevas variables obtenidas con Autoencoders en una arquitectura simple. Se observa que la combinación de un learning rate de 1 con el optimizador RMSprop es la que ofrece la solución más óptima (mayor diferencia de medianas).

La combinación que mejores resultados ofrece, es cuando se usa el optimizador RMSprop con un learning rate de 1. Se observa que prácticamente todas las variables presentan diferencias. Cuanto más fría es la estrella, más grandes son las diferencias. De nuevo, cuando las temperaturas de las estrellas están en el rango de 2600 - 2700 K las diferencias de las medianas son menores.

Una vez seleccionada la recombinación mas óptima se procede a ha hacer una visualización mas rigurosa con boxplot. En la figura 5.12 se muestran las nuevas componentes con detalle de cada uno de las rangos de temperaturas.

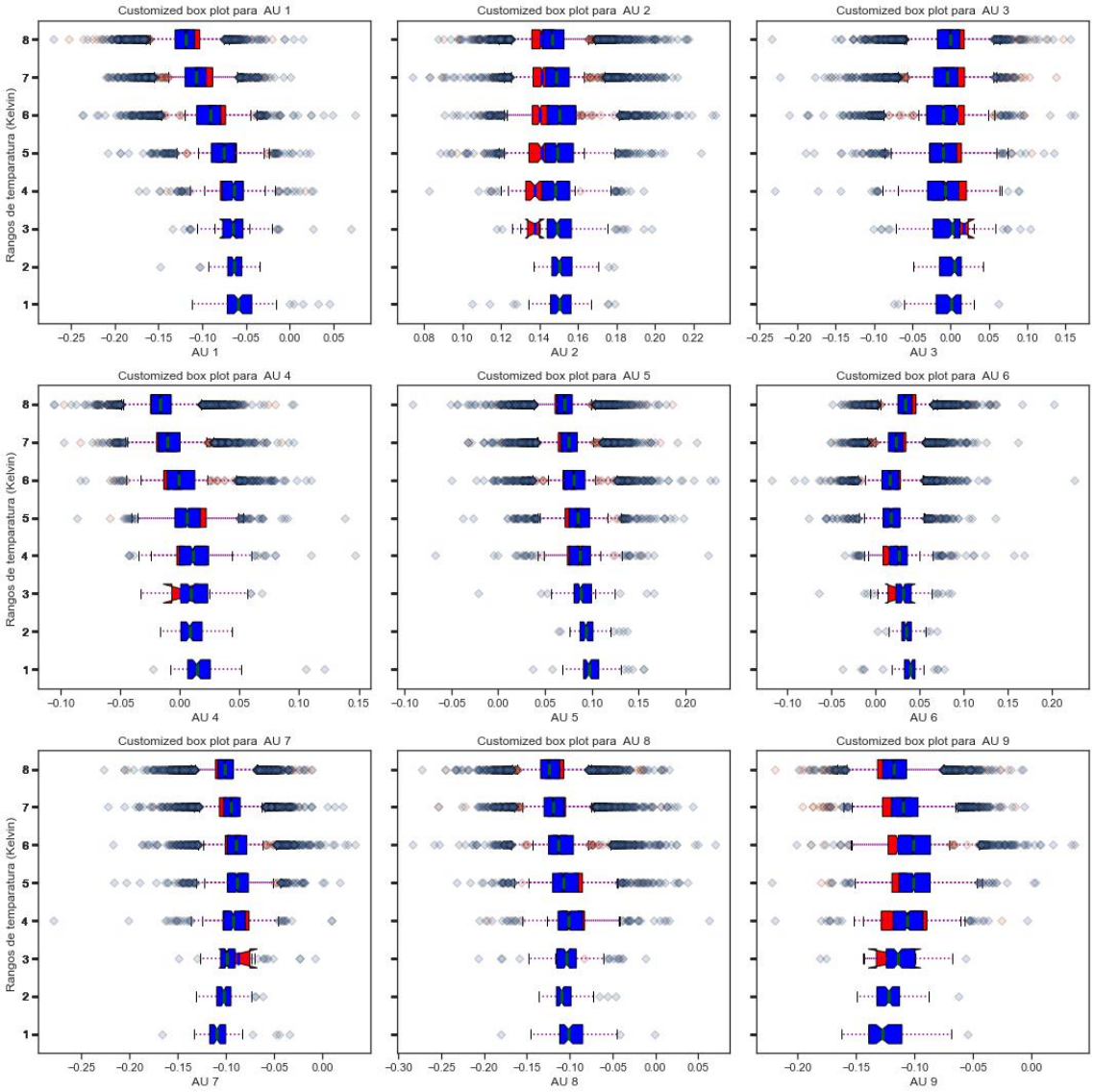

Figura 5.12: BoxPlot por rangos de temperatura de los valores de las observaciones en cada una de las nuevas variables obtenidas con Autoencoders de arquitectura simple. Se observa que prácticamente todas las nuevas variables presenta diferencias entre las medianas de los valores tomados por los tipos de estrellas. Cuando la temperatura de la estrella es menor, mayor diferencia presentan las medianas.

En la figura 5.13 se muestran los gráficos de dispersión de las variables que mayor diferencia de medianas han tenido. En concreto se han seleccionado las nuevas variables : $AU1$, $AU2$, $AU3$, $AU4$ y $AU5$

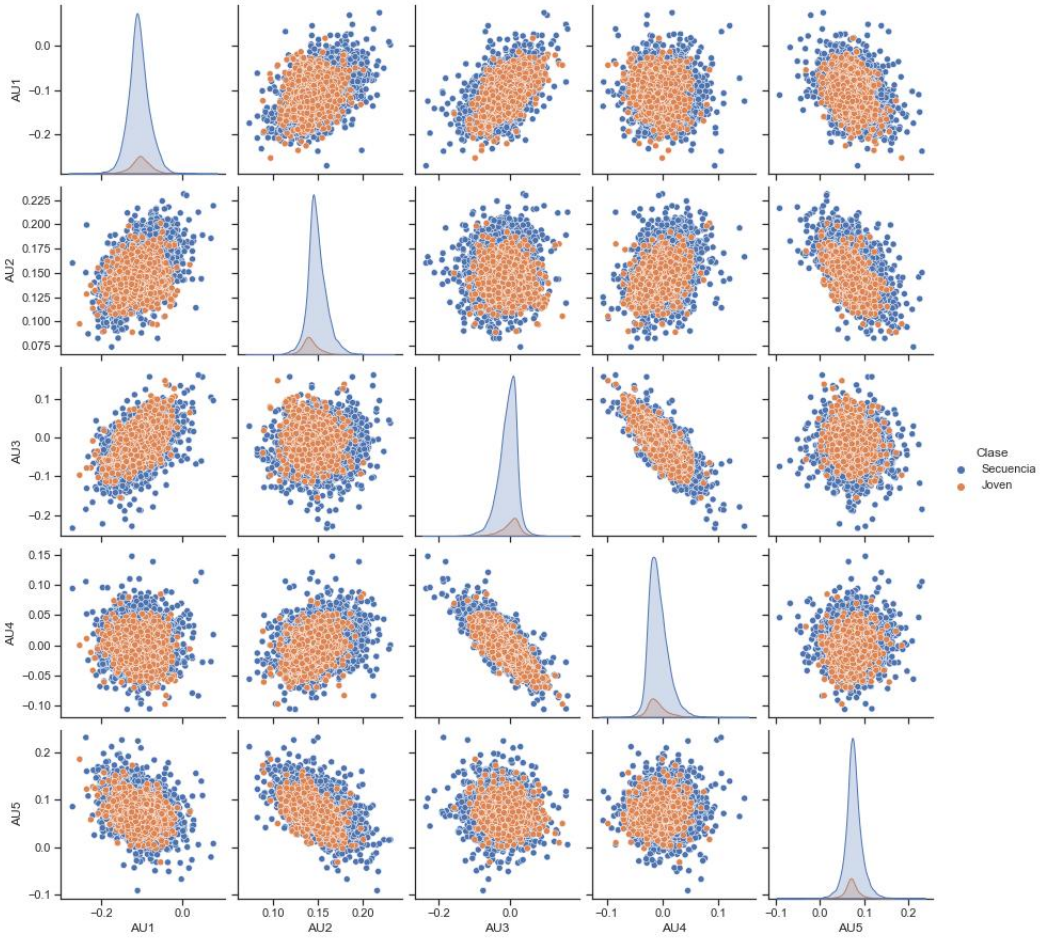

Figura 5.13: Representación en dos dimensiones de las observaciones en el nuevo espacio latente creado con Autoencoders de arquitectura simple. En naranja se muestran las estrellas jóvenes y en azul las estrellas en secuencia principal. En general, casi todas las estrellas (jóvenes y en secuencia principal) se localizan en una misma región.

En la figura 5.14 se muestra la representación gráfica de $AU4$ frente a $AU2$. Parece que estas dos nuevas variables son las que mejor separan en el nuevo espacio los espectros de las estrellas en secuencia principal de las jóvenes. No obstante, existe una zona con una alta acumulación de puntos, de donde no se puede distinguir el tipo de estrella.

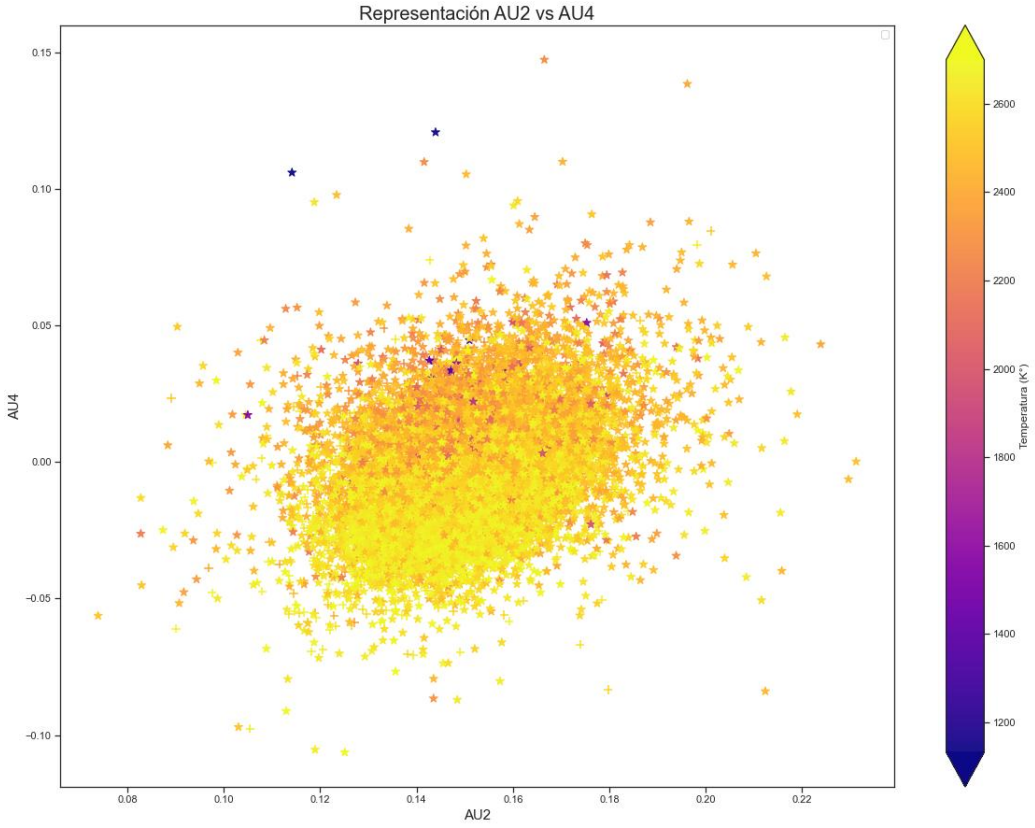

Figura 5.14: Representación gráfica de las nuevas variables $AU4$ vs. $AU2$ obtenidas con Autoencoders. Existe una zona donde se acumulan una gran mayoría de puntos y donde no se puede distinguir el tipo de estrella.

En la figura 5.15 se representa la densidad de probabilidad de ocurrencia de los valores $AU4$ y $AU2$ para los dos tipos de estrellas. Se puede observar como las zonas de alta densidad están muy cercanas, lo que produce un solape para un rango considerable de valores. La imagen no muestra claramente dónde se sitúa la densidad de probabilidad de ambos tipos de estrellas.

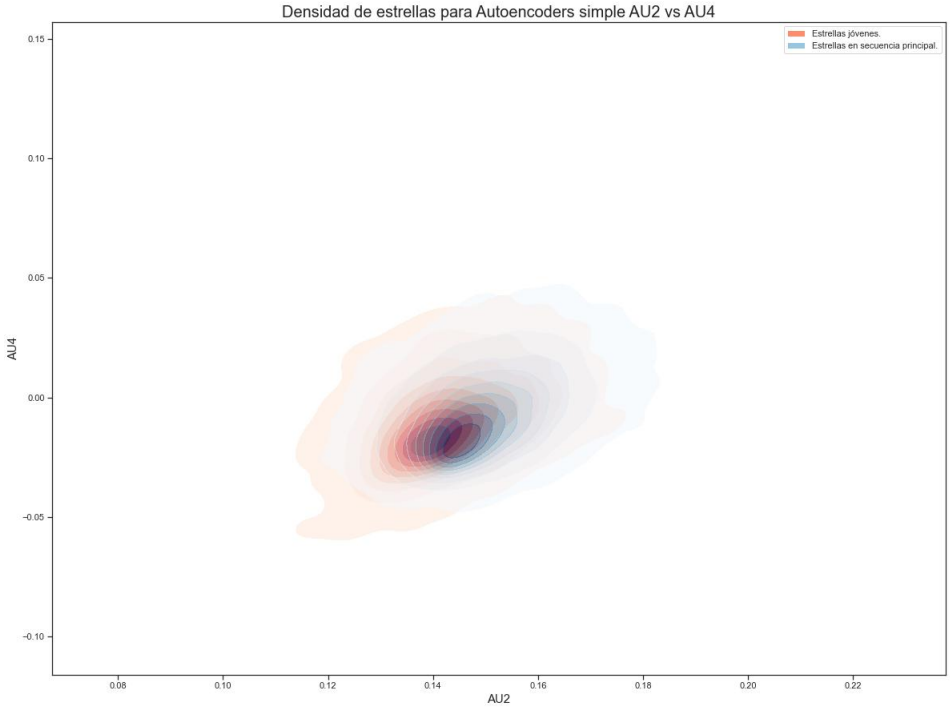

Figura 5.15: Representación gráfica de la distribución de densidad de probabilidad para los valores de $AU4$ y $AU2$ de las estrellas jóvenes y en secuencia principal. Esta densidad viene representada como una zona sombreada en color rojizo para las estrellas jóvenes y azulado para las estrellas en secuencia principal. Se puede observar que las zonas de alta densidad (zona saturada de colores) están muy próximas. No se muestra los valores de las densidades de probabilidad, ya que lo interesante no es en sí los valores, sino identificar las zonas de alta densidad que están ocupando los dos tipos de estrellas. Para estimar la densidad se ha usado Kernel Density Estimation que es un método no paramétrico que estima la función de densidad de variables aleatorias a partir de un número finito de observaciones, siendo esta función de densidad continua. En este caso la estimación se ha apoyado en un kernel con base Gaussiana.

5.3. Mapas de Difusión

En la tabla 5.3 se muestran listan las combinaciones utilizadas para la búsqueda de las mejores soluciones cuando se obtienen los nuevos espacios latentes con Mapas de Difusión. En total se ha obtenido 18 nuevos espacios, con sus correspondientes 9 nuevas variables ($MdD1$, $MdD2$,...$MdD9$). Las combinaciones se han obtenido en base a combinar los valores de σ , α y t.

La figura 5.16 muestra las diferencias entre la mediana de las estrellas jóvenes y la mediana de las estrellas en secuencia principal, para cada una de las nuevas variables.

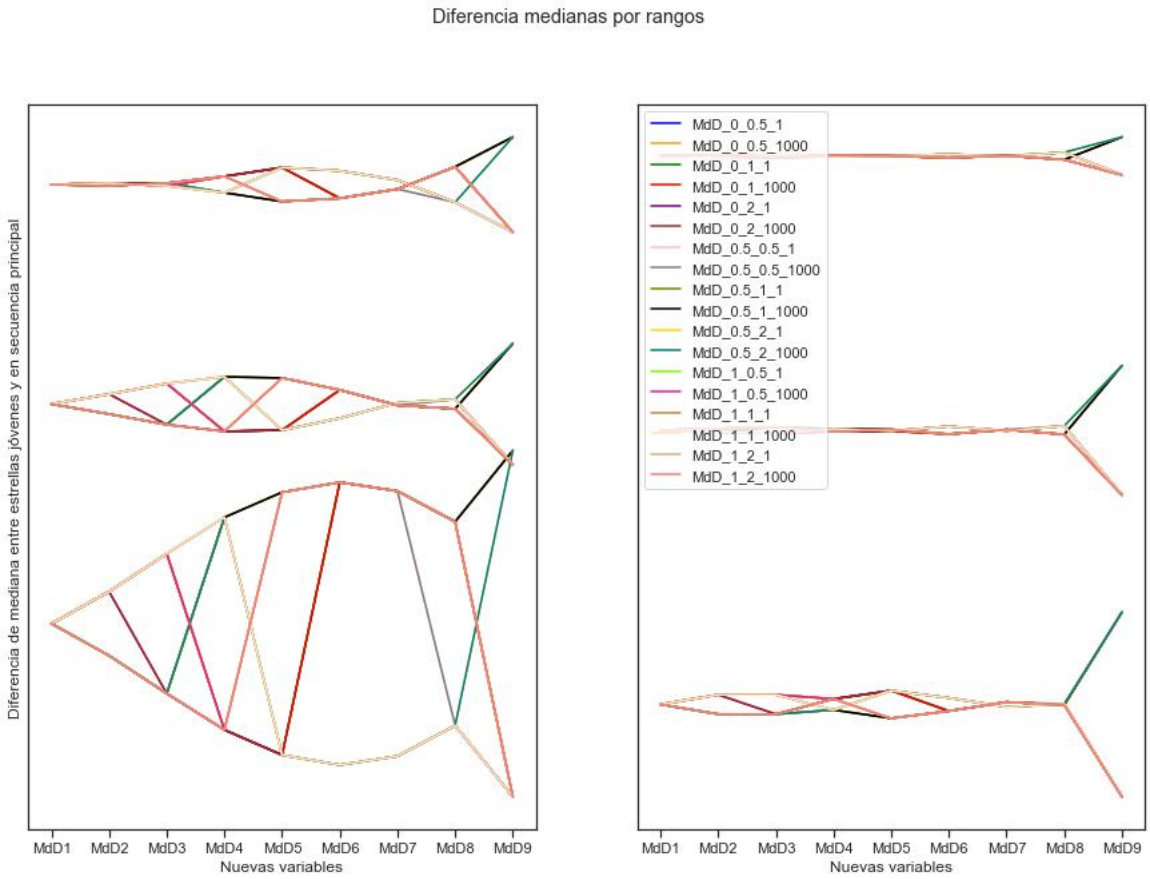

Figura 5.16: Representación gráfica de la diferencia de las medianas entre estrellas jóvenes y en secuencia principal en cada una de las nuevas variables obtenidas con Mapas de Difusión. Las diferencias entre las medianas se muestran por rangos de temperatura que abarcan 100 K comenzando en los 2000 K y terminando en 2700 K (a excepción del primer rango que abarca temperaturas inferiores a 2000 K).

Las combinaciones de α igual a 1 y 0.5 son las que mejores resultados están produciendo. En la figura 5.17 se muestran las diferencias para la combinación MdD_0_1_1000 y MdD_0.5_1_1000.

	α			
t	σ	0	0,5	1
1	0.5	MdD_0_0.5_1	MdD_0.5_0.5_1	MdD_1_0.5_1
1	1	MdD_0_1_1	MdD_0.5_1_1	MdD_1_1_1
1	2	MdD_0_2_1	MdD_0.5_2_1	MdD_1_2_1
1000	0.5	MdD_0_0.5_1000	MdD_0.5_0.5_1000	MdD_1_0.5_1000
1000	1	MdD_0_1_1000	MdD_0.5_1_1000	MdD_1_1_1000
1000	2	MdD_0_2_1000	MdD_0.5_2_1000	MdD_1_2_1000

Cuadro 5.3: Nombre de las combinaciones utilizadas para diferenciar entre las opciones de ejecución de Mapas de Difusión. Se concatena con "_" el acrónimo MdD, el valor de α, el valor de σ y el valor de t. Así por ejemplo, la combinación de $\alpha = 0,5$, $\sigma = 1$ y $t = 1000$ se llamará MdD_0.5_1_1000

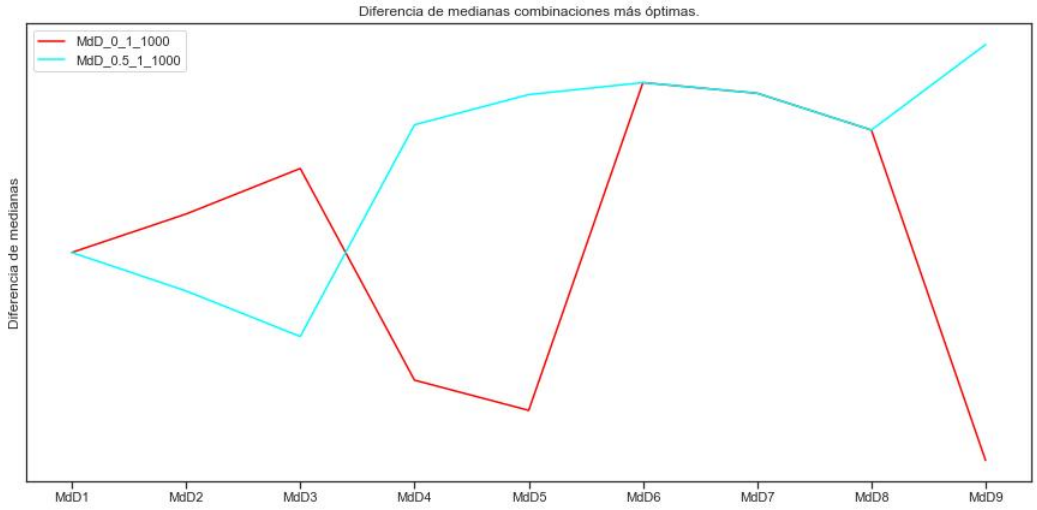

Figura 5.17: Representación gráfica de la diferencia de las medianas entre estrellas jóvenes y en secuencia principal en cada una de las nuevas variables obtenidas con Mapas de Difusión para valores de α de 0.5 y 1, σ igual a 1 y t igual a 1000. Se observa que prácticamente todas las variables presentan medianas diferentes paras los dos tipos de estrellas, siendo máximas las diferencias en las variables $MdD6$, $MdD8$ y $MdD9$. Se tomará la combinación de α igual a 0.5 como las más óptima.

La combinación que mejores resultados ofrece, es decir, que muestra mayores diferencias de medianas, es la que combina α igual a 0.5, σ igual a 1 y t igual a 1000. Prácticamente todas las nuevas variables están recogiendo los indicios de juventud de los espectros.

En la figura 5.18 se muestran el detalle para cada uno de los rangos de temperatura. Las estrellas jóvenes se muestran en rojo y las estrellas en secuencia principal en azul. Se observa que las nuevas variables $MdD8$, $MdD1$ y $MdD3$ separan mejor los tipos de estrellas, serán estas sobre las que se profundizará en las figuras siguientes.

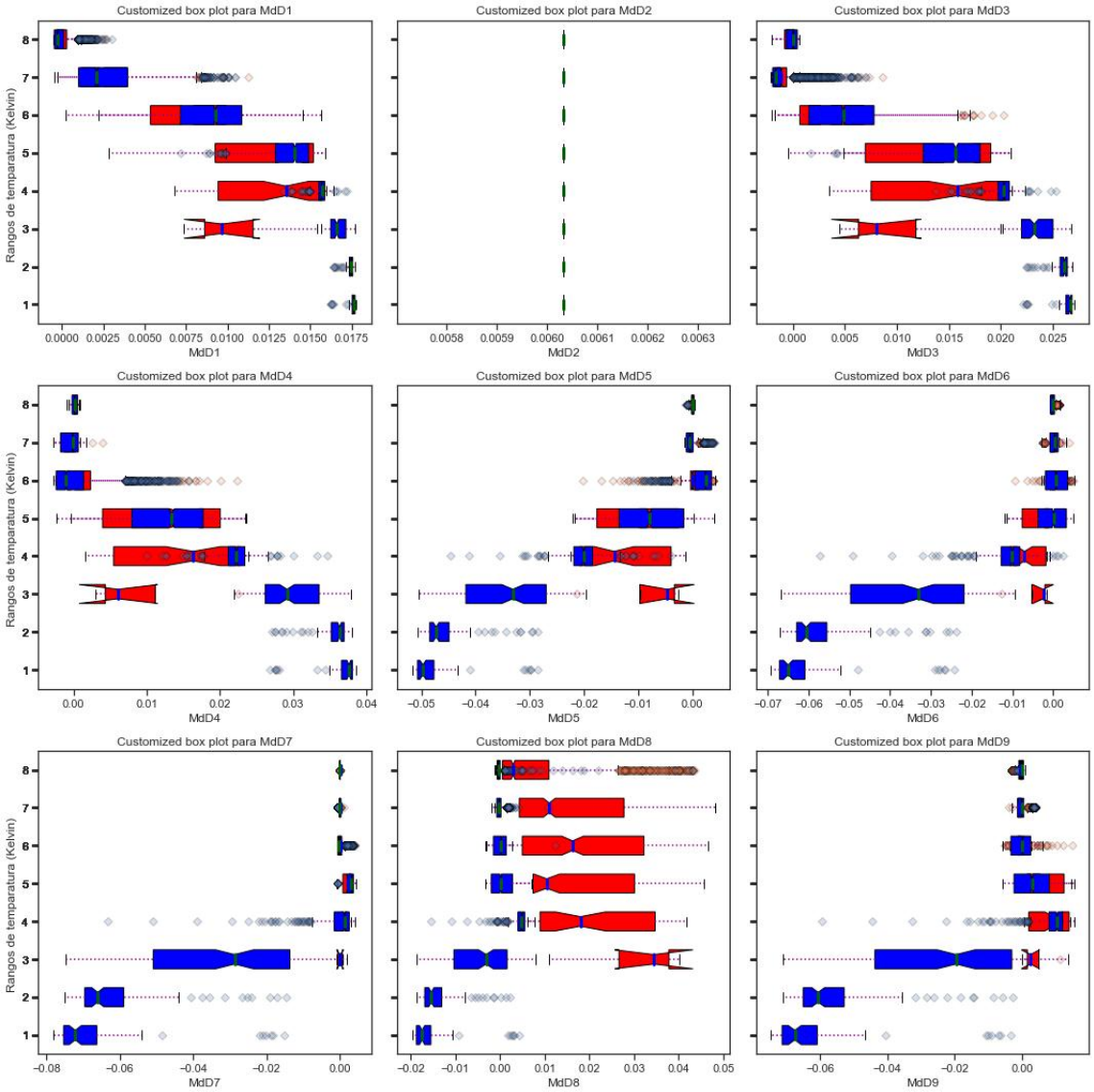

Figura 5.18: BoxPlot por rangos de temperatura de los valores de las observaciones en cada una de las nuevas variables obtenidas con Mapas de Difusión. Se observa que hay mayor separación de valores entre estrellas jóvenes (valores en rojo) y estrellas en secuencia principal (valores en azul) para la variable $MdD8$ en prácticamente todos los rangos de temperatura.

En la figura 5.19 se muestran los gráficos de dispersión de las observaciones en las nuevas variables que mejor han recogido los indicios de juventud, es decir en $MdD1$, $MdD6$, $MdD7$ $MdD8$ y $MdD9$. Se pueden localizar las estrellas en jóvenes dispersas por una amplia zona y separada de las estrellas en secuencia principal.

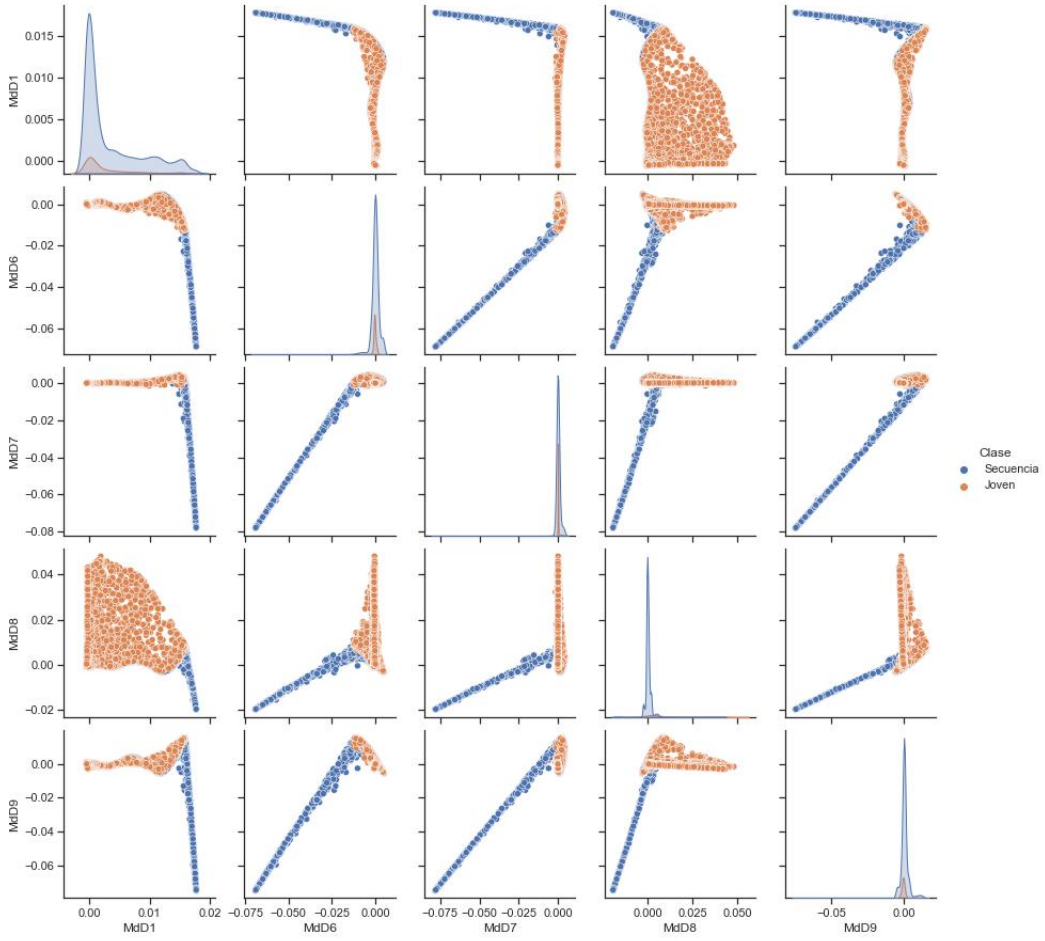

Figura 5.19: Representación en dos dimensiones de los valores de las observaciones en las nuevas variables obtenidas con Mapas de Difusión. En naranja se muestran las estrellas jóvenes y en azul las estrellas en secuencia principal. La representación bidimensional de $MdD1$ vs. $MdD8$ es la que muestra a las estrellas jóvenes ocupando una zona mayor y separada, por lo general, de la zona ocupada por las estrellas en secuencia principal. Será esta combinación sobre la que se profundizará en el trabajo para analizar cómo están comportándose estas variables en cuento a la edad y temperatura de las estrellas.

En la figura 5.20 se muestra la representación de $MdD1$ frente a la $MdD8$, por ser las dos nuevas variables que mejor consiguen separar las estrellas en secuencia principal de las estrellas jóvenes.

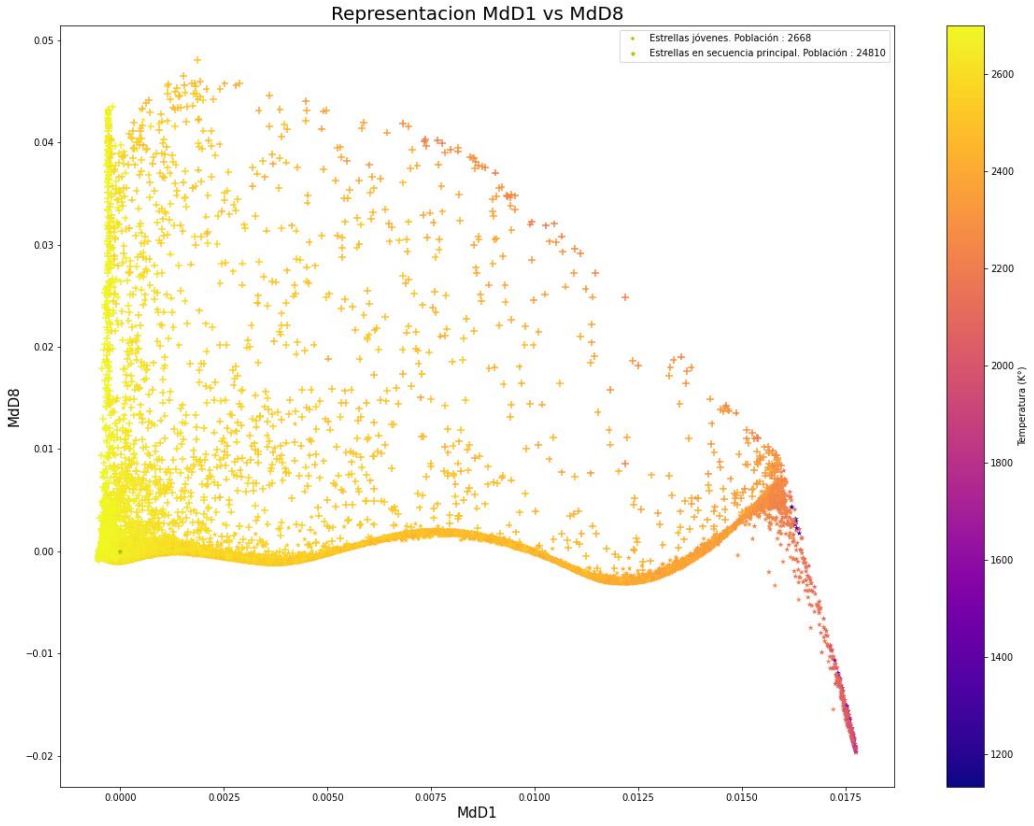

Figura 5.20: Representación gráfica de las observaciones en el espacio formado por las dos nuevas variables $MdD8$ vs. $MdD1$ obtenidas con Mapas de Difusión. Existe una zona pequeña, donde se acumulan una parte importante de los puntos, pero la mayoría de las estrellas se encuentran dispersas por toda la visualización. Se observa que los valores más grandes de $MdD1$ están reservados para estrellas en secuencia principal más frías y que los valores más grandes de $MdD8$ están reservados para las estrellas jóvenes.

En la figura 5.21 se representa la densidad de probabilidad de ocurrencia de los valores $MdD8$ y $MdD1$ para los dos tipos de estrellas. Se puede observar como las zonas de alta densidad están muy cercanas, lo que produce un solape para un rango considerable de valores, pero también que existe una zona grande de valores de $Md8$ para estrellas jóvenes que no tienen solape ninguno con las estrellas en secuencia principal. A través de las líneas de puntos se marcan las temperaturas medias de las zonas. En la parte inferior derecha (valores grandes de $MdD1$) se localizan las estrellas en secuencia principal frías, según va aumentando la temperatura los valores de $MdD1$ van decreciendo. Por otro lado, las estrellas jóvenes se localizan en valores mayores de $MdD8$ que las estrellas en secuencia principal. El comportamiento de la tendencia de los valores de $MdD1$ para las estrellas jóvenes es similar al de las estrellas en secuencia principal, a excepción de las estrellas que tienen temperaturas en el rango de 2100 - 2200 K.

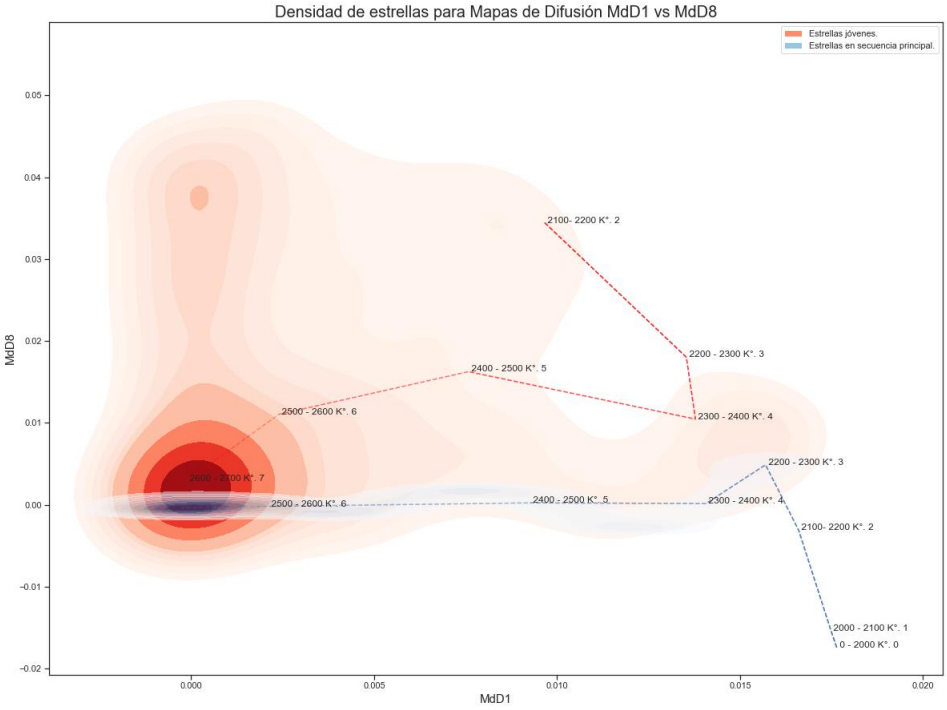

Figura 5.21: Representación gráfica de la distribución de densidad de probabilidad para los valores de $Md8$ y $MdD1$ de las estrellas jóvenes y en secuencia principal. Esta densidad viene representada como una zona sombreada en color rojizo para las estrellas jóvenes y azulado para las estrellas en secuencia principal. Se puede observar que las zonas de alta densidad (zona saturada de colores) están muy próximas. Se muestran las temperaturas medias de las estrellas que ocupan las diferentes zonas. Existe una tendencia de crecimiento de los valores de $MdD1$ según disminuye la temperatura de las estrellas para ambos tipos, a excepción de las estrellas jóvenes con temperaturas entre 2100 K y 2300K. No se muestra los valores de las densidades de probabilidad, ya que lo interesante no es en sí los valores, sino identificar las zonas de alta densidad que están ocupando los dos tipos de estrellas. Para estimar la densidad se ha usado Kernel Density Estimation que es un método no paramétrico que estima la función de densidad de variables aleatorias a partir de un número finito de observaciones, siendo esta función de densidad continua. En este caso la estimación se ha apoyado en un kernel con base Gaussiana.

Como se comentaba en el apartado 4.5, se ha contado con un conjunto extra de datos con los espectros de estrellas pertenecientes a diferentes zonas estelares. A modo de validación, se han localizado dichas estrellas en las representaciones bidimensionales de $MdD8$ vs. $MdD1$ para comprobar si las estrellas pertenecientes a zonas de formación estelar están situadas en la representación donde se espera, es decir, en valores de $MdD8$ más altos que los que tienen las estrellas en secuencia principal. Igualmente, se han localizado las estrellas de aquellas zonas que se saben no son jóvenes y que, por lo tanto, sus estrellas se pueden considerar como estrellas en secuencia principal, para validar que dichas estrellas ocupan la zona esperada. En la figura 5.22 se identifican las diferentes zonas estelares.

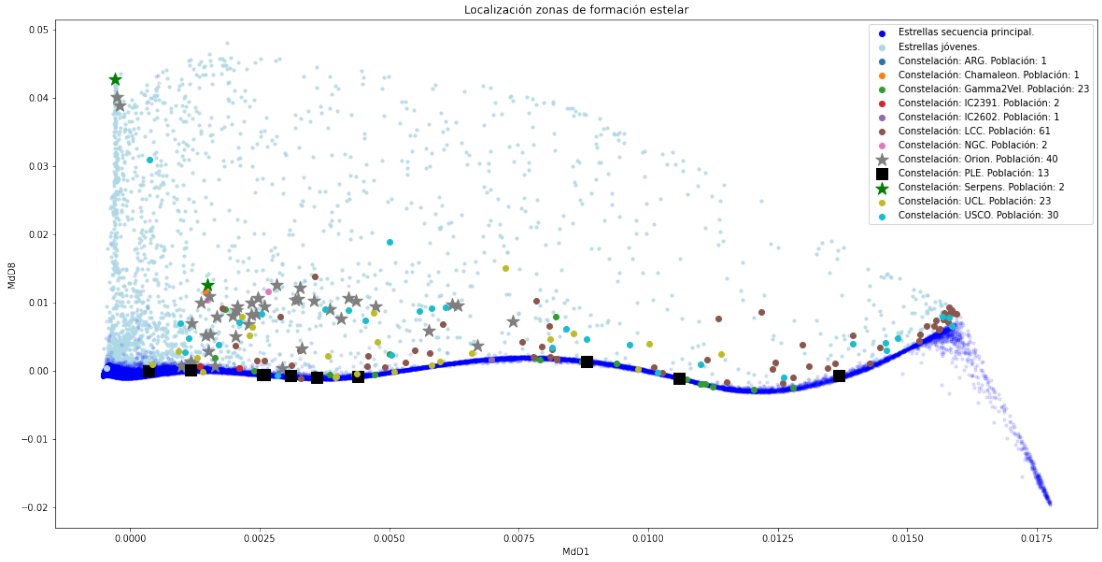

Figura 5.22: Representación gráfica de MD8 frente a MD1 para todo el conjunto de datos. Se localizan las estrellas que pertenecen a 12 zonas diferentes de formación estelar. Se sabe que Orion y Serpens son jóvenes y que PLE no es tan joven y que puede considerarse como pertenecientes a secuencia principal. Se observa que las zonas jóvenes se encuentran localizadas en los valores esperados para las estrellas jóvenes y que PLE se localiza en igualmente donde se espera a las estrellas que pertenecen a la secuencia principal.

Capítulo 6

Conclusiones

En este trabajo se ha experimentado con tres técnicas de reducción de dimensionalidad, comparándolas en como son capaces de captar y preservar los índicos de juventud contenidos en los espectros de estrellas ultrafrías.

En un análisis morfológico a alto nivel se puede observar una clara influencia de la edad y la temperatura en los espectros y, por lo tanto, puede analizarse construir herramientas que sean capaces de inferir dichas características a partir del espectro. Pero el espectro publicado por Gaia puede considerarse de alta dimensión, lo que puede suponer un problema a la hora de construir dichas herramientas. Son bien conocidas las bondades de las técnicas de reducción de dimensionalidad a la hora de afrontar este problema. Ante la gran variedad de técnicas existentes, se pretendía experimentar con Autoencoders y Mapas de Difusión, con el objetivo de analizar si en los nuevos espacios reducidos se recogían las características deseadas. Tomando como base los resultados obtenidos con PCA se puede concluir que los espacios de inferior dimensión creados con Mapas de Difusión captan mejor los indicios de juventud de los espectros, sin que eso suponga perdida de información sobre la temperatura.

PCA

Se ha comprobado que los nuevos espacios latentes de inferior dimensión creados con Principal Component Analysis captan la esencia de las diferencias entre los espectros de estrellas jóvenes y en secuencia principal de estrellas ultrafrías.

Las diferentes representaciones expuestas muestran como PCA diferencia los componentes principales de tal manera que separa un gran número de estrellas jóvenes de las estrellas en secuencia principal. Prácticamente, todos los enfrentamientos por pares de los componentes principales muestran para un conjunto considerable de estrellas valores diferenciados para los dos tipos de estrellas. Por lo tanto, se puede concluir que para ese grupo de estrellas PCA está encontrando indicios de juventud en los espectros publicados por Gaia. Pero no existe un par de componentes principales que separe todas las estrellas jóvenes de las estrellas en secuencia principal.

Tomando como ejemplo la representación del par de componentes $PCA4$ vs. $PCA3$ (figura 5.4) y añadiendo la temperatura de la estrella como una dimensión más (a través del color) se puede observar que las estrellas en secuencia principal tienen valores de la componente $PCA4$ más pequeños que las estrellas jóvenes, por lo que se puede

decir que la componente $PCA4$ contiene indicadores de la edad de las estrellas. No hay una separación clara de los tipos de estrella para los valores de las componentes. Siempre existe un grupo importante de estrellas jóvenes y en secuencia principal con temperaturas diferentes que se encuentran contenidas en una misma región, por lo que para esas estrellas los valores de $PCA3$ y $PCA4$ no son indicadores claros de la temperatura o la edad.

También se puede concluir que la componente $PCA3$ está recogiendo características de la temperatura de los espectros, ya que las estrellas en secuencia principal más frías tiene valores de la componente $PCA3$ más grandes que las estrellas calientes. Por lo tanto, no solo se están recogiendo indicios de la edad de las estrellas, sino que no se están perdiendo las características asociadas a la temperatura.

Un detalle de la representación de $PCA4$ frente a $PCA3$ que resulta de especial atención lo encontramos en el conjunto de estrellas en secuencia principal frías que presentan valores muy elevados de $PCA4$, cuando parece que los valores elevados de $PCA4$ están reservados para estrellas jóvenes. Este comportamiento aparece para 6 estrellas, el resto de las estrellas frías en secuencia principal se encuentra en la región esperada de $PCA3$.

Autoencoders

Se han probado dos arquitecturas diferentes de Autoencoders con varios tipos de optimizadores y varias tasas de aprendizaje y, se puede concluir, que ninguna de las combinaciones ha mejorado los resultados obtenidos por PCA.

Como sucedía con PCA, la creación de nuevos espacios latentes con Autoencoders parecen captar para un grupo de estrellas indicios de juventud en los espectros. En las representaciones expuestas se puede observar como algunas de las nuevas variables creadas consiguen separar los valores tomados por las estrellas jóvenes de los tomados por las estrellas en secuencia principal. Pero para esta técnica, los centros de densidad de probabilidad de ocurrencia de las observaciones para las nuevas variables son tan cercanos, que apenas se pueden diferenciar y, por lo tanto, no se pueden considerar como indicadores de juventud.

Mapas de Difusión

Los resultados obtenidos aplicando Mapas de Difusión son muy diferentes a los que se obtienen con las otras dos técnicas. En este caso se puede concluir que los nuevos espacios creados capturan muy bien los indicios de juventud de los espectros de estrellas ultrafrías publicados por Gaia.

Se puede observar una clara separación entre las estrellas jóvenes y las estrellas en secuencia principal para casi cualquier par de representaciones. Por ejemplo, la figura 6.1 muestra como la nueva variable $MdD8$ recoge claramente indicios de la edad. Las estrellas en secuencia principal se muestran valores pequeños de $MdD8$, siendo negativo para las estrellas muy frías, mientras que las estrellas jóvenes muestras valores de $MdD8$ más grandes. Es decir, la nueva variable $MdD8$ recoge los indicios de edad en los espectros. Por otro lado, la nueva variable $MdD1$ recoge claramente la influencia de la temperatura en los espectros, ya que las estrellas frías tienen valores de $MdD1$ grandes,

mientras que las estrellas más calientes tienen valores más pequeños.

La figura muestra por separado las coordenadas de difusión de las estrellas jóvenes y las estrellas en secuencia principal en el espacio bidimensional formado por las nuevas variables $MdD8$ y $MdD1$. Se puede ver en las dos figuras inferiores como las zonas ocupadas por los dos tipos de estrellas son muy diferenciados. En la parte superior de la figura no se diferencia por tipo de estrellas, dando más énfasis a la temperatura para comprobar como la variable $MdD1$ capta esta característica del espectro y, por lo tanto, no hay perdida de información relacionada con la temperatura.

Hay que comentar un par de detalles de las representaciones. El primero es como el grupo de estrellas en secuencia principal más frías tienen valores de $MdD8$ más pequeños que las más calientes y que se localizan en la parte inferior derecha en forma de apéndice pronunciado. Las más calientes se localizan al principio del apéndice (a la izquierda) y las más frías al final del apéndice (derecha), a excepción de un conjunto de 6 observaciones de estrellas más frías que no están al final del apéndice como deberían.

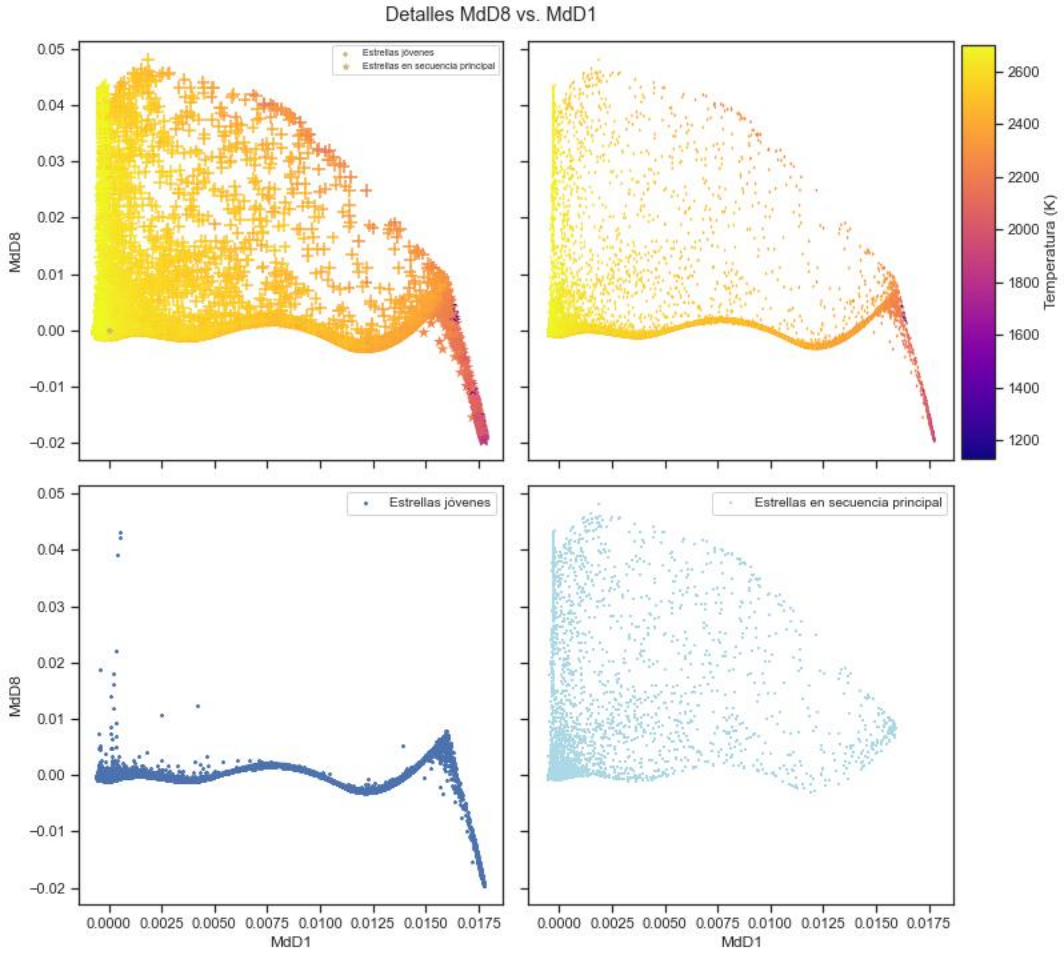

Figura 6.1: Representación con detalles de MdD8 vs. MdD1. En la parte superior izquierda se muestra la figura 5.20. En la parte superior derecha, se muestra la representación de Md8 frente a MdD1 usando mismo símbolo para las estrellas jóvenes y en secuencia principal, con el objetivo de visualizar mejor como MdD1 se ve afectado por la temperatura de las estrellas. En la parte inferior izquierda, se muestra únicamente las estrellas en secuencia principal para localizar más fácilmente el espacio que estás ocupan en la representación de MdD8 frente a MdD1. En la parte inferior derecha, se muestra únicamente las estrellas en jóvenes para localizar más fácilmente el espacio que estás ocupan en la representación de MdD8 frente a MdD1. Con estas dos últimas representaciones se puede ver como MdD8 se ve afectado por la edad de las estrellas.

El segundo detalle es sobre la temperatura de un conjunto de estrellas jóvenes que parece que no tienen valores para $MdD1$ como deberían tener (las estrellas más calientes tienen valores de $MdD1$ más pequeños y las más frías tienen valores más grandes).

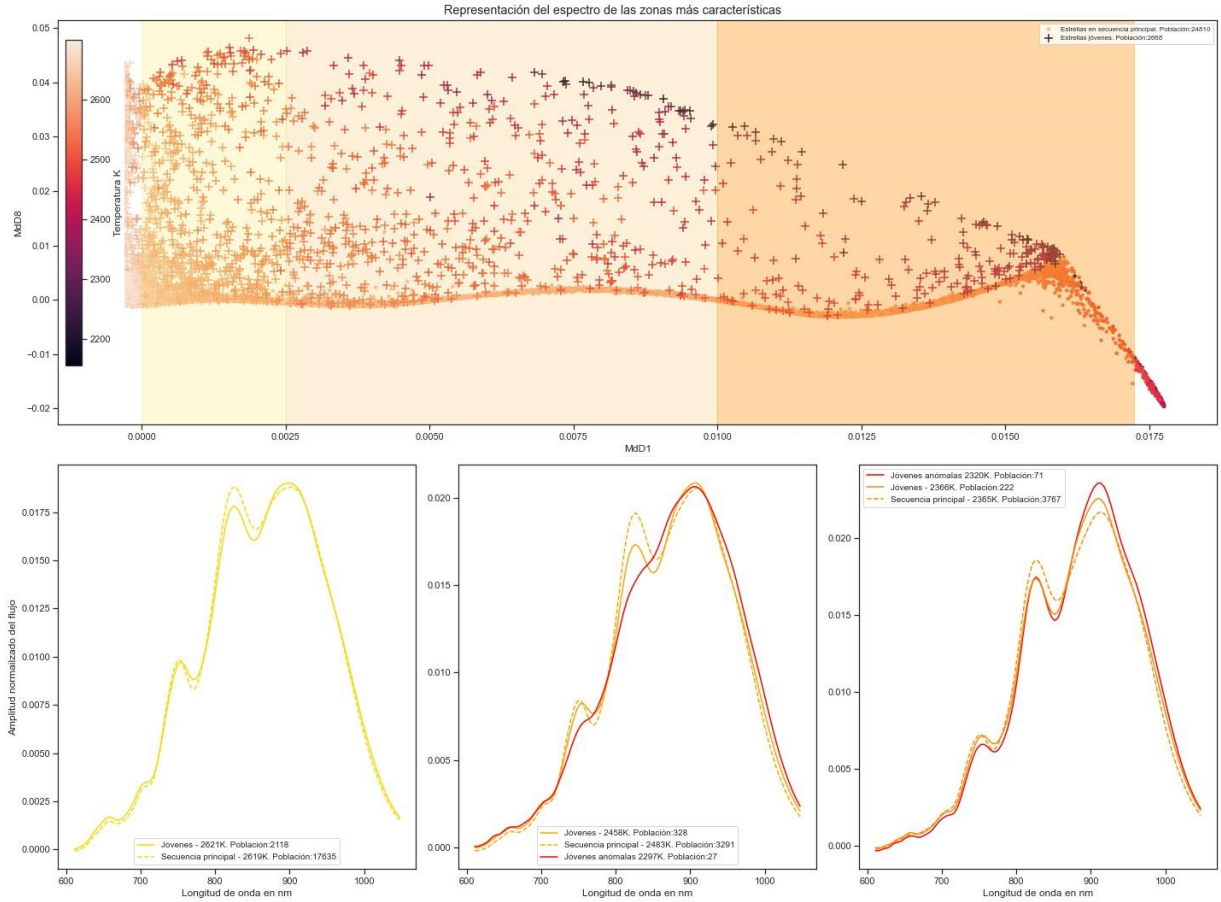

Figura 6.2: Representación con detalles de MdD8 vs. MdD1 diferenciando 5 zonas de temperatura de las estrellas. En la parte inferior de la representación, se muestran los espectros medios de las estrellas para cada una de las zonas.

En la figura 6.2 se muestran la media de los espectros de los dos tipos de estrellas agrupados en 3 zonas. Numerándolas de izquierda a derecha se tiene que:

- La primera zona contiene un conjunto de estrellas jóvenes con una temperatura media de 2510 K y un conjunto de estrellas en secuencia principal con temperatura media de 2599 K. Estas estrellas tienen altas temperaturas y son similares. Los espectros muestran diferencias mínimas, como se espera.

- La segunda zona contiene un conjunto de estrellas jóvenes con temperatura media de 2458 K y conjunto de estrellas en secuencia principal de 2483 K. En este caso las temperaturas siguen siendo altas y próximas. Se observa que las estrellas jóvenes situadas en el extremo superior tienen una temperatura baja. Comparando los tres tipos de espectros, se aprecia que dicho conjunto de estrellas tienen un espectro más plano, que es característico de las estrellas mucho más frías, mientras que los otros dos tipos de espectros siguen mostrándose muy cercanos.

- La tercera zona contiene un conjunto de estrellas jóvenes y en secuencia principal con temperatura media de 2365 K. Hay un conjunto de estrellas jóvenes con

temperatura media ligeramente inferior de 2320 K cuyo espectro medio muestra apenas diferencia.

Generales

Con las técnicas de creación de espacios latentes de baja dimensión se pueden hacer parametrizaciones más simples de los sistemas con poca pérdida de información.

En sistemas físicos complejos, la suposición de existencia de una relación lineal entre las observaciones de un conjunto de datos puede conducir a parametrizaciones de los sistemas subóptimas. Por esto, el uso de técnicas no lineales puede producir mejores resultados que las técnicas lineales. En este trabajo, se ha comprobado como con Mapas de Difusión se pueden crear nuevas variables que capturan mucho mejor indicios de edad y temperatura que la técnica lineal PCA.

Antes de decidir que técnicas son más apropiadas para este caso de uso, es interesante definir una forma de cuantificar lo bien que estás sirven al propósito específico. Este trabajo se ha decantado por cuantificarlas en función de la diferencia de medianas en las nuevas variables para cada tipo de estrellas. Para otros casos de uso se puede tomar como referencia cuánto es el error que se obtiene al usar las nuevas variables obtenidas para hacer una inferencia sobre la edad de la estrella en función del número de variables tomadas. O también, por ejemplo, cuánto de buenas son las nuevas variables a la hora de reconstruir el espacio original, en este caso, a la hora de reconstruir los espectros originales recogidos por Gaia. Puede quedarse esta reflexión para trabajos futuros.

Trabajos futuros

Tanto con PCA, Autoencoders y Mapas de Difusión se observa un comportamiento anómalo para un conjunto de 6 observaciones de estrellas de muy baja temperatura. Sería interesante analizar esos espectros en detalle y compararlo con espectros del mismo tipo de estrellas para intentar profundizar en el motivo de la anomalía. En el apéndice A Se incluyen los identificadores de dichas estrellas y sus espectros.

Apéndice A

Espectros anómalos

En la tabla A.1, se listan los identificadores proporcionados en DR3 de Gaia, de las seis estrellas que muestran un comportamiento anómalo a la hora de obtener sus coordenadas de difusión a través de Mapas de Difusión. También se muestran sus espectros en la figura A.1, comparándolos con las medias de los espectros de estrellas con rango de temperatura cercanos. Todas estas estrellas se encuentra en el ciclo de vida llamado secuencia principal.

Source ID
2997171394834174976
3426333598021539840
1037131492704550656
1267906854386665088
5052876333365036928
4752399493622045696

Cuadro A.1: Identificador de la fuente proporcionado por Gaia en DR3 de espectros anómalos por la posición que ocupan sus datos en los nuevos espacios latentes.temperatura.

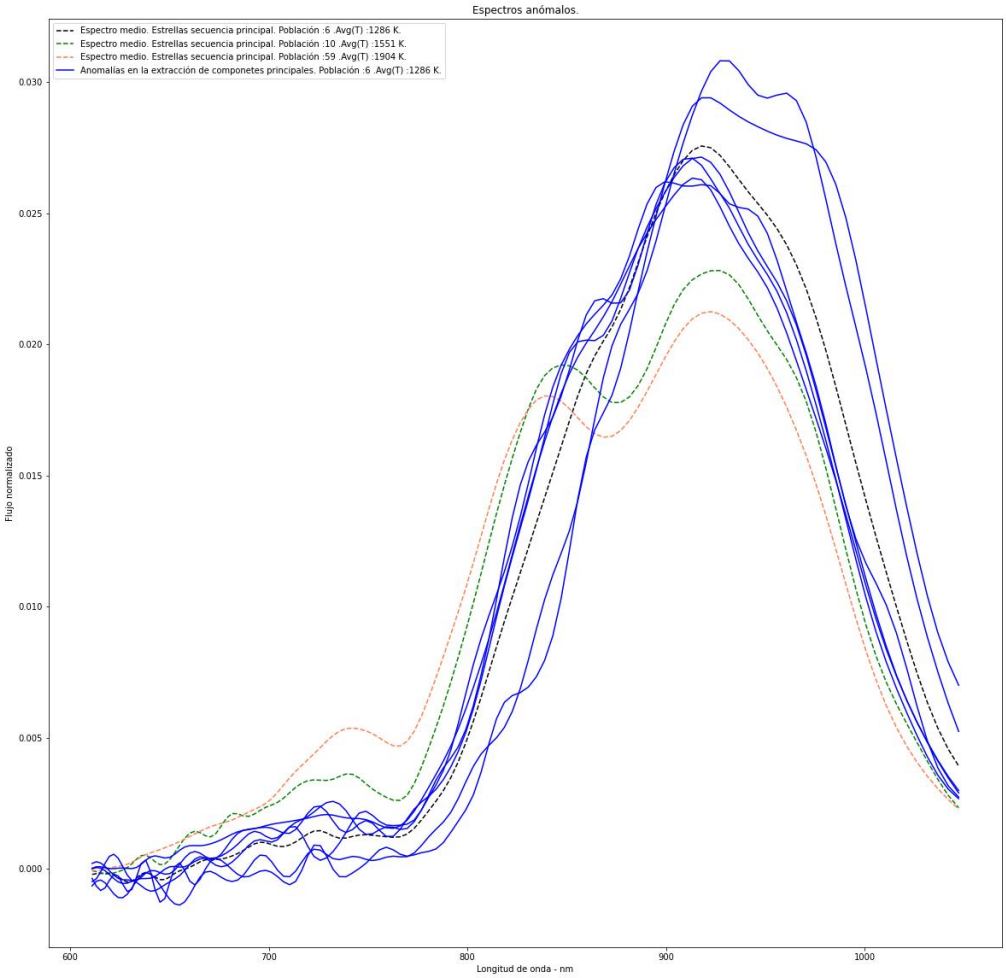

Figura A.1: Espectros anómalos por la posición que ocupan sus datos en los nuevos espacios latentes. Estas estrellas pertenecen a las consideradas en secuencia principal y son las 6 que menos temperatura tienen. Sus temperaturas son: 1132 K, 1147 K, 1354 K, 1356 K, 1361 K y 1365 K

Bibliografía

[AALV+21] Benjamin Ades-Aron, Gregory Lemberskiy, Jelle Veraart, John Golfinos, Els Fieremans, Dmitry S. Novikov, and Timothy Shepherd. Improved task-based functional mri language mapping in patients with brain tumors through marchenko-pastur principal component analysis denoising. *Radiology*, 2021.

[BTB22] P. Bacon, A. Trovato, and M. Bejger. Denoising gravitational-wave signals from binary black holes with dilated convolutional autoencoder. 2022.

[Col16] Gaia Collaboration. The gaia mission. *TBD*, nov 2016.

[CPG05] Adams F. C., AU Bodenheimer P., and Laughlin G. M dwarfs: planet formation and long term evolution, 2005.

[CR05] Lee AB Coifman RR, Lafon S. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. 2005.

[DCD+17] Allen B. Davis, Jessi Cisewski, Xavier Dumusque, Debra A. Fischer, and Eric B. Ford. Insights on the spectral signatures of stellar activity and planets from PCA. *The Astrophysical Journal*, 846(1):59, aug 2017.

[DH85] Hinton GE Sejnowski TJ Ackley DH. A learning algorithm for boltzmann machines. 1985.

[ESA22a] ESA. https://www.cosmos.esa.int/web/gaia/dr3. *TBD*, 2022.

[ESA22b] ESA. https://www.esa.int/. *TBD*, 2022.

[GGBG19] Amin Ghafourian, Orestis Georgiou, Edmund Barter, and Thilo Gross. Wireless localization with diffusion maps, 2019.

[GWMMG20] D. M. Gaslac Gallardo, S. M. Giuliatti Winter, G. Madeira, and M. A. Muñoz-Gutiérrez. Analysing the region of the rings and small satellites of neptune. *Astrophysics and Space Science*, 365(1), jan 2020.

[Gé19] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition.* O´Reilly Media,Inc., 2019.

[Hot33] H Hotelling. Analysis of a complex of statistical variables into principal. 1933.

[JDKI97] T. J. Henry J. D. Kirkpatrick and M. J. Irwin. Ultra-cool m dwarfs discovered by qso surveys.i: The apm objects. 1997.

[JKR+22] Anirudh Jonnalagadda, Shubham P. Kulkarni, Akash Rodhiya, Hemanth Kolla, and Konduri Aditya. A co-kurtosis based dimensionality reduction method for combustion datasets, 2022.

[LD22] L T Lehmann and J-F Donati. Diagnosing large-scale stellar magnetic fields using PCA on spectropolarimetric data. *Monthly Notices of the Royal Astronomical Society*, 514(2):2333–2345, jun 2022.

[MGPLW21] Marco A. Muñoz-Gutiérrez, Antonio Peimbert, Matthew J. Lehner, and Shiang-Yu Wang. Long-term dynamical stability in the outer solar system. i. the regular and chaotic evolution of the 34 largest trans-neptunian objects. *The Astronomical Journal*, 162(4):164, sep 2021.

[MP67] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices., 1967.

[MVS+21] Eric A. Moreno, Jean-Roch Vlimant, Maria Spiropulu, Bartlomiej Borzyszkowski, and Maurizio Pierini. Source-agnostic gravitational-wave detection with recurrent autoencoders, 2021.

[Pea01] K Pearson. On lines and planes of closest fit to systems of points in space. 1901.

[PPVC21] Stephen K. N. Portillo, John K. Parejko, Jorge R. Vergara, and Andrew J. Connolly. Dimensionality reduction of SDSS spectra with variational autoencoders, 2021.

[PTC+19] Zachary J. Pace, Christy Tremonti, Yanmei Chen, Adam L. Schaefer, Matthew A. Bershady, Kyle B. Westfall, Mé déric Boquien, Kate Rowlands, Brett Andrews, Joel R. Brownstein, Niv Drory, and David Wake. Resolved and integrated stellar masses in the SDSS-iv/MaNGA survey. i. PCA spectral fitting and stellar mass-to-light ratio estimates. *The Astrophysical Journal*, 883(1):82, sep 2019.

[RBH+21] Alexey Ryabov, Bernd Blasius, Helmut Hillebrand, Irina Olenina, and Thilo Gross. Estimation of functional diversity and species traits from ecological monitoring data, 2021.

[Sho16] Seth Shostak. New search for signals from 20,000 star systems begins, 2016.

[slhls] scikit-learn https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

[SP12] Joseph W. Richards Darren Homrighausen Peter E. Freeman Chad M. Schafer and Dovi Poznanski. Semi-supervised learning for photometric supernova classification. *Royal Atronomy Society*, 2012.

[TJI+22] Kurihana Takuya, Franke James, Foster Ian, Wang Ziwei, and Moyer Elisabeth. Insight into cloud processes from unsupervised classification with a rotationally invariant autoencoder, 2022.

[WZZ+22] Xuetong Wang, Kanhao Zhao, Rong Zhou, Alex Leow, Ricardo Osorio, Yu Zhang, and Lifang He. Normative modeling via conditional variational autoencoder and adversarial learning to identify brain dysfunction in alzheimer's disease. 2022.